UNIVERSITY OF CALIFORNIA,
IRVINE

Extrinsic Regulation of Mammalian mRNA 3’ Processing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Liang Liu

Dissertation Committee:
Professor Yongsheng Shi, Chair
Professor Rozanne M. Sandri-Goldin
Professor Bert L. Semler
Professor Klemens J. Hertel

2023

DEDICATION

To

My father

My mother

and

My grandmother

TABLE OF CONTENTS

Page

LIST OF FIGURES iv

LIST OF TABLES vi

ACKNOWLEDGEMENTS vii

VITA ix

ABSTRACT OF THE DISSERTATION xiii

CHAPTER 1: Introduction 1

1.1 Overview of the mRNA 3' processing complex 2

1.2 Mechanism and regulation of alternative polyadenylation 15

1.3 Modulation of mRNA 3' processing during viral infection 19

1.4 mRNA 3' processing as an attractive drug target 24

1.5 Summary 26

CHAPTER 2: Mechanism and consequence of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation 27

2.1 Publication Note 27

2.2 Summary 28

2.3 Introduction 29

2.4 Results 31

2.5 Discussion 52

2.6 Methods 59

CHAPTER 3: The anti-cancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3' processing machinery 64

3.1 Summary 65

3.2 Introduction 66

3.3 Results 68

3.4 Discussion 102

3.5 Methods 119

CHAPTER 4: Prospectives 132

REFERENCES 135

LIST OF FIGURES

Page

Figure 1.1 Cis elements of mRNA 3' processing 4

Figure 1.2 Mammalian mRNA 3' processing complex 6

Figure 1.3 Alternative polyadenylation 18

Figure 2.1 HSV-1 infection induces widespread and dynamic APA changes 32

Figure 2.2 Normalized PAS-seq read counts at proximal and distal PAS of APA genes 34

Figure 2.3 HSV-1-induced APA changes occur, at least in part, co-transcriptionally 36

Figure 2.4 4sU-seq signal ratio of the intronic PAS of DtoP genes 38

Figure 2.5 ICP27-dependent and -independent HSV-1-induced APA changes 40

Figure 2.6 Significant APA changes between KOS and ΔICP27 virus-infected cells 41

Figure 2.7 Mechanism of HSV-1-mediated APA regulation 44

Figure 2.8 Significant APA changes comparison in cells under different conditions 45

Figure 2.9 HSV-1-mediated APA regulation and mRNA export 48

Figure 2.10 DoTT transcripts are processed at cryptic PAS 49

Figure 2.11 HSV-1-induced intronically polyadenylated APA isoforms are translated 53

Figure 2.12 Ribo-seq tracks for example genes 54

Figure 2.13 A model for HSV-1-mediated APA regulation 56

Figure 3.1 Compound 2-mediated inhibition of mRNA 3′ processing in vitro is sequence-dependent 70

Figure 3.2 Compound 2 does not affect 3′ processing complex assembly on resistant RNA 72

Figure 3.3 Cleavage site (CS) region is a major determinant of Compound 2 sensitivity 74

Figure 3.4 In vitro cleavage for L3 and SVL chimeras 76

Figure 3.5 Massively parallel in vitro assay (MPIVA) for charactering the relationship between CS sequence and drug sensitivity 80

Figure 3.6 Mutating the YA cleavage site does not affect PAS sensitivity to

Compound 2 82

Figure 3.7 Characterization of SVL variants by MPIVA 83

Figure 3.8 Additional validated CS variants from both backbones from in vitro MPIVA 84

Figure 3.9 6-mer motif analyses of top 10,000 resistant and sensitive CS variants from both backbones from the in vitro MPIVA 86

Figure 3.10 Machine learning analysis of Compound 2 sensitivity 88

Figure 3.11 C3PO layer 1 learned sequence features 90

Figure 3.12 C3PO layer 2 learned sequence features 91

Figure 3.13 ΔG of minimum free energy structures and base pairing probabilities of the top 10,000 resistant and sensitive sequences 93

Figure 3.14 JTE-607 modulates poly(A) site selection in a sequence-specific manner 97

Figure 3.15 JTE-607 modulates transcription termination in a sequence-specific manner 98

Figure 3.16 Conservation of JTE-607-sensitive and –resistant PASs 101

Figure 3.17 A model for sequence-specific inhibition pf pre-mRNA 3′ processing by Compound 2 104

LIST OF TABLES

Page

Table 3.1 Activity and IC_{50} of all PASs tested in this study 108

Table 3.2 R2 for various machine learning architectures tested 109

Table 3.3 Log (12.5 μM/DMSO) and C3PO predicted score for all PASs 111

Table 3.4 Sequence of all in vitro tested PAS in this study 112

Table 3.5 Oligonucleotides and synthesized DNA used in this study 116

ACKNOWLEDGEMENTS

First, I would like to thank my family: my mom, dad, and grandma: for raising me up, for giving me the opportunity to study abroad in the United States, for encouraging me to pursue what I truly enjoy, and for always being there for me during the up and down times. Ten years away from home in a different country alone is not easy, and I certainly would not be where I am today without your infinite love and support.

I would like to thank my high school biology teacher Dr. Chuan Ni for stimulating my initial interest in biology. I'm honored to be the very first student in your teaching career to pursue a biology Ph.D. I would also like to thank Dr. Rajeev Misra, my undergraduate PI for teaching me all the lab tricks, helping me to apply for my very first undergraduate research grant, and introducing me to the Microbiology graduate student association so I got exposed to how graduate school feels like. I also want to thank Mellecha, Jae, Brenda, Julian and Nan for all your mentorship, friendship, and of course all the Friday nights fun times with beer.

Next, I would like to thank my Ph.D. advisor Dr. Yongsheng Shi for allowing me to explore virology projects in the lab, teaching me how to think critically and be a good scientist, encouraging me to present my research at conferences, and being very open to talk about pretty much anything with me, from project big picture to specific techniques to career development. Without your high standard and perfectionism, I would not be who I am today. I would also like to thank you for your accurate judgement on what types of projects are suitable for different people in the lab. If it is not you, I would never know I really like doing in vitro assays.

Now it is the time for my labmates. Xiuye was the first person I talked to when I started rotating in the lab and ended up being the person that I talk to the most. Thank you for always being patient with me when I have a million questions to ask you. Also, thank you for pausing your own experiments to help my experiments many times because you are worried about my several days of work getting wasted. Almost all the techniques I know today were learned from you, and I truly appreciate your mentorship. Kristianna, you have been a wonderful friend since I started my rotation. You are always so caring, supportive, and you always check to make sure I am okay when you notice I am stressed. Thank you for helping me grow scientifically and at the same time forcing me to have a work life balance. It was also nice to do the Friday nights' dinner and Trader Joe's runs with you and occasionally working together on the weekend. I really miss that. Lindsey, we join the lab around the same time and I really admire your independence on driving a project forward. I also enjoy our scientific discussions and collaborations. Your critical thinking and presentation skills are something I always want to learn from and improve on. Marielle, although you are the youngest graduate student in lab, I really like chatting with you. Your curiosity and positive energy always make me feel the same too. Also, I'm glad to have some accompany during the late nights in the lab. Yoseop, we are similar in a way that we both have very high standard on experiments and we always pick the method that gives highest quality. I enjoy collaborating with you and I believe you and I can achieve very high-level goals in the future. I admire your critical thinking and your ability to pick up one totally unfamiliar project and carry it forward. Also, thank you for working together with me during the shift working late nights. Those are some of my hardest times but I am glad I was not alone in lab. Lusong, you are relatively new to the lab but you are always willing to offer help on my experiments, which I am really grateful. Nabila and Yong were also important people that helped me a lot in my early days in the lab. I feel extremely luck that everyone in the lab gets alone very well, willing to help each other and the lab has a very collaborative atmosphere.

I would like to thank members of my thesis committee. Dr. Rozanne Sandri-Goldin, thank you for your support and help on my virology project. Your tremendous amount of knowledge on herpes virus is something I truly admire and I would love to be someone like you in my own field of research in the future. Dr. Klemens Hertel, thank you for the hard questions you asked me during my talks. You encouraged me to think deeper about my project and consider the areas that I am not always familiar with. Dr. Bert Semler, you are my committee member and also my first-year advisor. You are the person I feel most comfortable to talk to apart from my PI. Thank you for your valuable advice on both research projects and my personal matters. You said to me during our very first meeting six years ago that your goal as an advisor is to make sure I do well and succeed. Now that six years has passed, I hope I am able to make you proud and help you achieve your goal.

The MMG department is the best department I have ever been in. The department has such a warm and friendly environment. Everyone knows each other and is willing to help each other. I would like to thank the office staff Janet, Shanti, and Lesley for helping me taking care of my international student related problems when HR is not responding to me. I would also like to thank the following people in the department: Michele, Alexis, MyPhuong, Hung, Jessie A., Francisco, and José, who I frequently go to when I have questions or need help with.

Last but certainly not least, this section is dedicated to a few special friends of mine. Runlong Guo, we are friends for ten years now. Thank you for visiting me in Irvine multiple times during my hardest time in graduate school. I honestly don’t think I can survive my first year without your support and friendship. Nina Pham and Jasmine Ho, I am grateful that we became good friends after I finished TAing for Anatomy. Thank you for introducing me to your little group so I met Cintia, Crystal, Cherry, and Bo. I really enjoyed our times together trying authentic Vietnams’ food and thank you for inviting me to your New Year’s parties. The note you wrote for me lives on my desk in lab now. Tessa Chou, we only know each other for less than a year but I feel we are similar in many ways. I really enjoy chatting about and trying different foods with you. It is always very peaceful and relaxing when we eat out together, some feeling that I don’t have when I eat out with others. I also really admire your clear goals and dedication on funding and future career. If you are reading this one day: Keep trying and don’t give up! You can do it! I believe in you!

I thank the fundings from NIH R21 AI166703 and UCI CVR graduate fellowship provided by the Graduate Division. Chapter 2 of this dissertation is a reprint of the material as it appears in: Xiuye Wang, Liang Liu, Adam W Whisnant, Thomas Hennig, Lara Djakovic, Nabila Haque, Cindy Bach, Rozanne Sandri-Goldin, Florian Erhard, Caroline C. Friedel, Lars Dölken, Yongsheng Shi. (2021) Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation. PLoS Genet 17(3): e1009263. https://doi.org/10.1371/journal.pgen.1009263, used with permission from PLoS Genetics under the Creative Commons license.

VITA

Liang Liu

EDUCATION

Doctor of Philosophy in Biomedical Sciences 2023
University of California, Irvine, CA

Bachelor of Science in Microbiology (Summa Cum Laude) 2017
Arizona State University, Tempe, AZ

RESEARCH EXPERIENCE

Graduate Research Assistant 2017-2023
University of California, Irvine, CA

Undergraduate Research Assistant 2015-2017
Arizona State University, Tempe, AZ

PUBLICATIONS

* Denote equal contribution

1. **Liu, L.***, Yu, M A.*, Wang, X., Soles, L., Chen, Y., Yoon, Y., Sarkan, K., Valdez, M.C., Linder, J., Yu, Z., Qiao, F., Li, W., Seelig, G., Shi, Y. (2023) The anti-cancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3' processing machinery. ***Nat. Struct. Mol. Biol.*** In revision

2. Wang, X.*, **Liu, L.***, Whisnant, A.W., Hennig, T., Djakovic, L., Haque, N., Bach, C., Sandri-Goldin, R.M., Erhard, F., Friedel, C.C., Dölken, L., Shi, Y. (2021) Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation. ***PLoS Genet*** 17(3): e1009263.

FELLOWSHIP AND AWARDS

Center for Virus Research Graduate Fellowship 2020-2023
University of California, Irvine, CA

School of Life Sciences Undergraduate Research Fellowship 2016-2017
Arizona State University, Tempe, AZ

New American University Scholarship 2016-2017
Arizona State University, Tempe, AZ

SELECTED ORAL/POSTER PRESENTATIONS

Oral Presentations:

HIV mRNA 3' Processing: Mechanism, regulation, and therapy 2023
(NIAID, NIH Virology Training Grant Symposium 2023, UC Irvine)

Regulation of mammalian mRNA metabolism by viral infection 2022
(NIAID, NIH Virology Training Grant Symposium 2022, UC Irvine)

mRNA 3' processing regulation by the anti-cancer compound JTE-607 2022
(Department of Microbiology & Molecular Genetics Seminar, UC Irvine)

Mechanistic dissection of the anti-cancer compound JTE-607-mediated inhibition of mRNA 3' processing 2021
(Cold Spring Harbor Laboratory 2021 Eukaryotic mRNA Processing Meeting, Virtual)

Virus-mediated regulation of mRNA 3' processing 2021
(NIAID, NIH Virology Training Grant Symposium 2021, Virtual, UC Irvine)

Mechanism of mRNA 3' processing inhibition by the anti-cancer compound JTE-607 2021
(Department of Microbiology & Molecular Genetics Seminar, UC Irvine)

Poster Presentations:

Sequence-dependent regulation of mRNA 3' processing by the anti-cancer compound JTE-607 2022

(The 6th Annual UCI School of Medicine Grad Day, UC Irvine)

The anti-cancer compound JTE-607-mediated inhibition of pre-mRNA 3' processing is sequence-dependent — 2022
(RNA 2022: the 27th Annual RNA Society Meeting, Boulder, Colorado)

Mechanism of herpes simplex virus 1-mediated regulation of host mRNA 3' processing and alternative polyadenylation — 2022
(Center for Virus Research 2022 Retreat, UC Irvine)

Mechanistic dissection of the anti-cancer compound JTE-607-mediated inhibition of mRNA 3' processing — 2021
(The 5th Annual UCI School of Medicine Grad Day, UC Irvine)

Loss of inter-protomer interactions overcomes the drug binding pocket defect of the AcrB protein of *Escherichia coli* — 2017
(American Society for Microbiology, The 56th Annual Arizona and Southern Nevada Branch Meeting, U Arizona)

Functional characterization of *Escherichia coli* multi-drug efflux pump protein AcrB — 2017
(The 24th Annual Undergraduate Research Poster Symposium, ASU)

TEACHING EXPERIENCE

Teaching assistant and guest lecturer — 2019
Bio D170 Applied Human Anatomy
University of California, Irvine, CA

Teaching assistant — 2019
Bio 93 DNA to Organisms
University of California, Irvine, CA

LEADERSHIP AND SERVICE

Departmental Graduate Student Representative — 2022-2023
Department of Microbiology & Molecular Genetics, School of Medicine
University of California, Irvine, CA

Departmental Ultracentrifuge Coordinator and Trainer 2020-2023
Department of Microbiology & Molecular Genetics, School of Medicine
University of California, Irvine, CA

Departmental Seminar Search Committee 2022-2023
Department of Microbiology & Molecular Genetics, School of Medicine
University of California, Irvine, CA

Departmental Seminar Search Committee 2020-2021
Department of Microbiology & Molecular Genetics, School of Medicine
University of California, Irvine, CA

OUTREACH

Executive Board and Science Fair Committee Member 2019-2020
ReachOut TeachOut
University of California, Irvine, CA

ABSTRACT OF THE DISSERTATION

Extrinsic Regulation of Mammalian mRNA 3' Processing

by

Liang Liu
Doctor of Philosophy in Biomedical Sciences
University of California, Irvine, 2023
Professor Yongsheng Shi, Chair

Eukaryotic mRNA 3' processing is an essential step in the gene expression pathway and it is tightly coupled to many other cellular processes including transcription termination, splicing, mRNA export, and translation. As a result, it is under stringent regulation by many different factors such as viral infection, stress, and small molecule inhibitors. This study examines how mRNA 3' processing and transcription termination are regulated during herpes simplex virus 1 (HSV-1) infection and during treatment of JTE-607, a small molecule inhibitor. We found that HSV-1 and its immediate early factor ICP27 contribute to widespread changes in host mRNA alternative polyadenylation (APA) by activating intronic polyadenylation sites (PAS). We also discovered that the small molecule inhibitor JTE-607 inhibits mRNA 3' processing in a sequence-specific manner and the cleavage site region immediately following AAUAAA is the core sensitivity determinant of a PAS to JTE-607. A machine learning model that predicts the impact of JTE-607 on PAS selection and transcription termination genome-wide was developed. The findings of this study provide insights into the regulatory mechanism of mRNA 3' processing and transcription termination by viral infection and small molecule inhibitors, which shed light on targeting mRNA 3' processing as a novel strategy for broad spectrum anti-viral and anti-cancer therapeutics.

CHAPTER 1:

Introduction

The 3' ends of the vast majority of eukaryotic mRNAs undergo mRNA 3' processing, a two-step process characterized by the cleavage and polyadenylation of the 3' end of pre-messenger RNA (pre-mRNA) (Colgan and Manley 1997; Chan et al. 2011). mRNA 3' processing is an essential step in eukaryotic mRNA biogenesis, and it directly impacts other processes within the gene expression pathway, such as splicing, transcription termination, mRNA export, translation, and mRNA turnover (Colgan and Manley 1997; Zhao et al. 1999; Moore and Proudfoot 2009). In addition to being an essential step in mRNA biogenesis, mRNA 3' processing is also important in gene expression regulation. In humans, over 70% of genes contain more than one polyadenylation site (PAS) and thus can be cleaved and polyadenylated at multiple alternative PASs. This is called alternative polyadenylation (APA). APA generates mRNA isoforms that either contain the same coding region but different 3′ UTR or mRNAs with different coding regions, both of which can be differentially regulated and in turn can affect mRNA stability, translation efficiency, and mRNA/protein localization (Shi and Manley 2015; Tian and Manley 2017; Chan et al. 2011; Soles and Shi 2021).

mRNA 3′ processing and APA are regulation hotspots for different factors including but not limited to viral infection, stress and small molecule inhibitors (Nemeroff et al. 1998; Rutkowski et al. 2015; Jia et al. 2017; Hennig et al. 2018; Wang et al. 2020, 2021; Ross et al.

2020). Dysregulation of mRNA 3′ processing and APA has also been associated with different diseases such as cancer and neurological disorders (Masamha et al. 2014; Lukiw and Bazan 1997). As a result, it is essential to identify the proteins involved in mRNA 3′ processing, evaluate their functions, and understand their regulation to combat different diseases or viral infections and facilitate the development of novel therapeutic strategies potentially using mRNA 3′ processing as a druggable node. This chapter will provide an overview of the components of mammalian mRNA 3′ processing and its mechanism of regulation, specifically during viral infection and by a small molecule inhibitor.

1.1 OVERVIEW OF THE mRNA 3’ PROCESSING COMPLEX

Assembly of the mRNA 3′ processing complex is initiated by co-transcriptional recruitment and binding of various trans-acting factors (proteins) to their cis elements (RNA sequences). This involves a series of RNA-protein and protein-protein interactions. Some of these interactions are weak by themselves but when different complexes assemble on the pre-mRNA, they form a stable 3′ processing complex. In this section, the core RNA and protein components of mRNA 3′ processing will be described.

Cis Elements:

Multiple short RNA sequences at the pre-mRNA 3′ end collectively form the polyadenylation site (PAS) (Figure 1.1). There are two types of PASs: canonical and non-canonical. The most conserved cis element with in an canonical PAS is the hexamer sequence A(A/U)UAAA, which is present in about 70~80% PASs (Beaudoing et al. 2000; MacDonald and Redondo 2002; Chan et al. 2011). Deletion or mutation of this core hexamer sequence significantly decreases cleavage efficiency of the pre-mRNA, indicating the importance of this sequence in mRNA 3′ processing. The actual cleavage position, the position at which pre-mRNA is cleaved by the 3′ processing machinery, is usually immediately after a CA or UA dinucleotide located 10~30 nucleotides downstream of the core A(A/U)UAAA hexamer (Chen et al. 1995). Although CA or UA are the two cleavage sites most frequently used, human PASs are very diverse and pre-mRNA can also be cleaved after other dinucleotide sequences such as GA, CG etc. (Bogard et al. 2019). Finally, approximately 30 nucleotides further downstream of the cleavage position is a region called downstream elements (DSE). The DSE is a stretch of U/GU-rich sequences, which is more variable and not found in all human PASs (Chan et al. 2011). Together, the core A(A/U)UAAA hexamer, the C/UA cleavage position, and the U/GU-rich DSE determine the general location of a PAS.

In addition to the hexamer, cleavage position, and the DSE, many PASs also contain other auxiliary sequences located either upstream or downstream of the A(A/U)UAAA hexamer. These auxiliary sequences serve as binding sites for different trans-acting factors in the 3′ processing complex and, in turn, can influence the efficiency of mRNA 3′ processing at specific PASs. Upstream auxiliary sequences are usually U-rich whereas auxiliary sequences downstream are generally G-rich (Chan et al. 2011). Some PASs also contain one or more UGUA motif(s) that are enriched approximately 50 nucleotides upstream of the cleavage position (Zhu et al. 2018).

About 20~30% of PASs are non-canonical and do not contain the A(A/U)UAAA

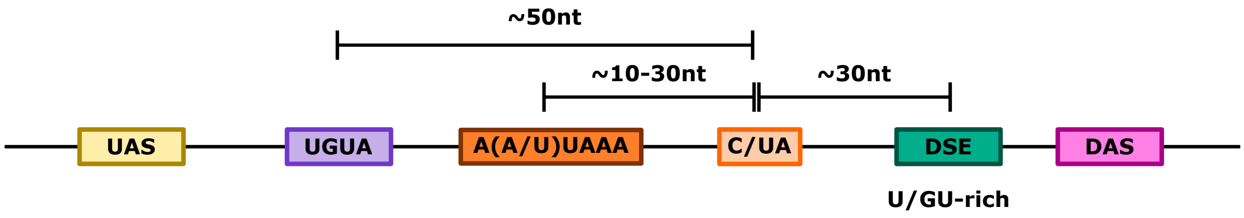

Figure 1.1 Cis elements of mRNA 3′ Processing

Schematic representation of the cis elements for mammalian mRNA 3’ processing

UAS: upstream auxiliary sequences

DSE: downstream elements

DAS: downstream auxiliary sequences

hexamer sequence (MacDonald and Redondo 2002; Beaudoing et al. 2000). In these PASs, additional auxiliary sequences may aid in the recruitment of 3′ processing factors. The precise mechanism for non-canonical PAS regulation remains unclear and will not be further discussed here.

Trans-Acting Factors:

The trans-acting factors of mammalian mRNA 3′ processing complex have been identified and characterized in detail through an RNA affinity pull down and mass spectrometry approach (Shi et al. 2009). Remarkably, there are over 85 proteins present in the purified 3′ processing complex. These include known core protein complexes involved in mRNA 3′ processing as well as other proteins that may mediate crosstalk of 3′ processing with other processes such as transcription, splicing, and translation. The core mammalian mRNA 3′ processing complex consists of four multi-subunit protein complexes: CPSF (**C**leavage and **P**olyadenylation **S**pecificity **F**actor), CstF (**C**leavage **St**imulatory **F**actor), and CFIm and CFIIm (**C**leavage **F**actor I and II) (Mandel et al. 2007; Shi et al. 2009; Tian and Manley 2017). Other proteins that are involved in mRNA 3′ processing include RNA polymerase II (RNAPII), polyA polymerase (PAP), and Rbbp6. CFIIm and Rbbp6 were both present in the 3′ processing complex originally purified by RNA affinity pull down, although their interactions with the other 3′ processing factors and RNA are weak and are not considered to be stably associated factors. The importance of CFIIm and Rbbp6 in mRNA 3’ processing was uncovered recently through the successful in vitro reconstitution of cleavage and polyadenylation. CPSF, CstF, CFIIm, and Rbbp6 are required for cleavage in vitro, whereas only CPSF is believed to be required for

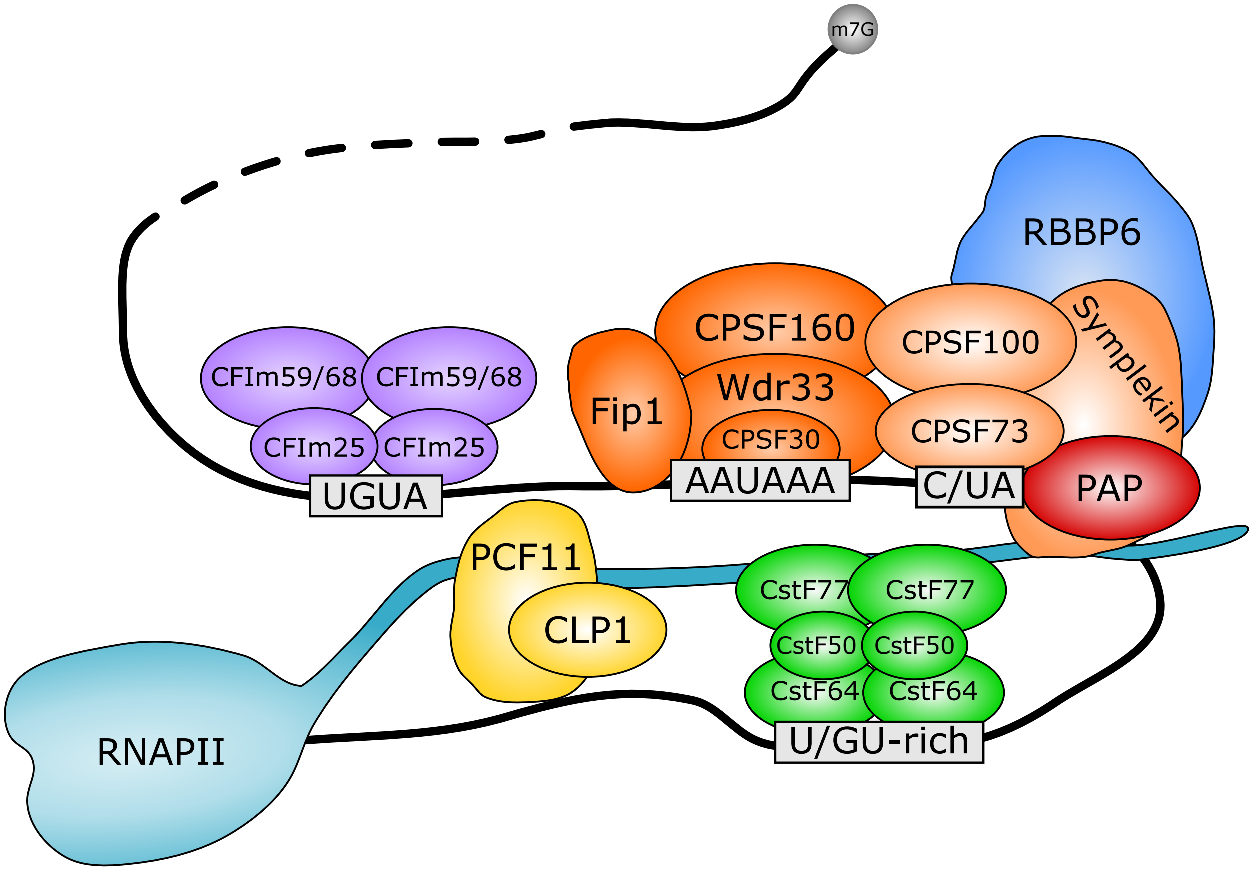

Figure 1.2 mammalian mRNA 3' Processing Complex

Schematic representation of the mammalian mRNA 3' processing complex

CPSF: cleavage and polyadenylation specificity factor

CstF: cleavage stimulatory factor

CFIm: cleavage factor I

CFIIm: cleavage factor II

RNAPII: RNA polymerase II

polyadenylation (Schmidt et al. 2022; Boreikaite et al. 2022). In this section, I will describe each of the protein complexes in detail, including their function and interaction partners within the 3′ processing complex (Figure 1.2).

CPSF:

The cleavage and polyadenylation specificity factor (CPSF) consists of 7 subunits, namely CPSF30, CPSF73, CPSF100, CPSF160, Fip1, Wdr33, and Symplekin (Shi et al. 2009). CPSF is essential for both cleavage and polyadenylation as it binds to the pre-mRNA at A(A/U)UAAA hexamer and cleaves RNA after a C/A or U/A motif downstream of the core hexamer. CPSF can be further divided into two sub-complexes: the polyadenylation specificity factor (mPSF) and the cleavage factor (mCF). mPSF includes CPSF30, CPSF160, Wdr33, and Fip1 and is important for RNA binding; whereas mCF is comprised of CPSF73, CPSF100, and Symplekin and is involved in the endonucleolytic cleavage step during mRNA 3′ processing. In this section, I will discuss the role of each protein within the two subcomplexes.

mPSF:

The mammalian polyadenylation specificity factor (mPSF) is responsible for RNA binding. mPSF recognizes the A(A/U)UAAA hexamer with high specificity as a single U to G mutation at the third nucleotide abolished mRNA 3′ processing (Bienroth et al. 1991). Early studies suggested that CPSF160 and CPSF30 are involved in A(A/U)UAAA recognition. First, two proteins of ~160 kDa and ~30 kDa in the purified CPSF complex (thought to correspond to CPSF160 and CPSF30, respectively) can be specifically cross-linked to AAUAAA-containing RNA substrates (Bienroth et al. 1991; Moore et al. 1988). Furthermore, in a pull down assay, recombinant human CPSF160 preferentially bound to AAUAAA-containing RNA (Murthy and Manley 1995). However, a more recent study in 2014 demonstrated that Wdr33 (which has similar size as CPSF160 at ~160 kDa) and CPSF30 are in fact the two proteins that directly bind to the A(A/U)UAAA hexamer (Chan et al. 2014). This finding was further confirmed in 2018 using a cryo-EM structural approach that determined the structure of CPSF160, Wdr333, and CPSF30 bound to an AAUAAA-containing RNA (Sun et al. 2018). Within the AAUAAA hexamer, A1 and A2 are recognized by zinc finger 2 of CPSF30 and A4, A5 are bound by zinc finger 3 of CPSF30. The U3 and A6 nucleotides form a Hoogsteen base pair and interact with Wdr33. CPSF160 serves as a scaffold protein that does not directly bind to AAUAAA RNA but instead recruits Wdr33 and CPSF30 to the correct position for PAS recognition.

The final protein of mPSF is Fip1, which binds to U-rich sequences upstream of A(A/U)UAAA using its arginine-rich C-terminus (Kaufmann et al. 2004). In addition to its RNA binding property, Fip1 is also an essential factor that mediates protein-protein interactions between CPSF and other proteins or protein complexes during mRNA 3' processing. For example, Fip1 interacts with CFIm through its RE/D domain and by doing so, enhances the usage of certain PASs, contributing to APA (Lackford et al. 2014). Fip1 also recruits PAP to the cleaved RNA to facilitate polyadenylation (Murthy and Manley 1995; Kaufmann et al. 2004).

mCF:

The mammalian cleavage factor (mCF) consists of 3 proteins: CPSF73, CPSF100, and Symplekin. CPSF73 and CPSF100 share similar domain structure. Both proteins contain a metallo-β-lactamase domain and a β-CASP domain and belong to the β-CASP subfamily of the zinc-dependent metallo-β-lactamase proteins, which contains a number of known DNA and RNA endonucleases (Callebaut et al. 2002; Mandel et al. 2006). Despite these similarities, the two proteins are currently believed to have different roles in regulating 3’ processing activity. CPSF73 is the active endonuclease within the mRNA 3′ processing complex. CPSF100 is thought to lack nuclease activity because several key zinc-binding residues have been mutated in the *S. cerevisiae* homolog Cft2p (Aravind 1999). However, more detailed sequence analysis shows that these residues are present and highly conserved in CPSF100 homologues in other species, from the fission yeast *S. pombe* to humans. This raises the possibility that CPSF100 is an active endonuclease in most species and *S. cerevisiae* is an outlier. However, direct biochemical evidence on CPSF100 nuclease activity is still lacking, and it remains unclear if CPSF100 is an active nuclease within the 3′ processing complex.

The third protein within mCF is Symplekin, whose function is still not fully understood but it is generally thought to be a scaffold protein responsible for the recruitment of other mRNA 3′ processing factors and stabilization of the entire complex. It also interacts with CstF, and disruption of Symplekin-CstF interaction negatively affects mRNA 3′ processing (Ruepp et al. 2011).

CstF:

The cleavage stimulatory factor (CstF) consists of three proteins: CstF77, CstF64, and CstF50. Each of these proteins exists as a dimer in the CstF complex (Takagaki and Manley 1997). CstF64 also contains a closely related paralog, CstF64tau. The CstF complex binds to downstream U/GU-rich elements within the RNA and is required for pre-mRNA cleavage but not polyadenylation (Mandel et al. 2007; Chan et al. 2011; Boreikaite et al. 2022).

CstF binding to downstream elements is achieved by the RNA recognition motif (RRM) of CstF64, which can bind to U/GU-rich sequences similar to downstream elements independently of the rest of the poly(A) machinery (Takagaki and Manley 1997). Purified CstF complex binds to U/GU-rich RNA sequences weakly as shown by UV-crosslinking experiments. However, this binding can be enhanced by adding increasing concentrations of purified CPSF into the reaction suggests that CstF and CPSF cooperatively bind to A(A/U)UAAA containing RNAs (Murthy and Manley 1992). Such a mechanism may be important to ensure full complex assembly and efficient 3′ processing only on RNAs that contain both the A(A/U)UAAA hexamer and a U/GU-rich downstream elements. It should be noted that unlike CPSF, which has a very well-defined binding site A(A/U)UAAA, the CstF binding site is more variable and variable sites are bound with different affinities. This may suggest a role for CstF in alternative polyadenylation (Yao et al. 2012). CstF64 has a closely related paralog called CstF64tau (CstF64τ), which is highly expressed in the testis. Studies on mice have shown that CstF64τ knock out causes defects in spermatogenesis and infertility in males (Dass et al. 2007). CstF64τ has a similar domain structure as CstF64, and both paralogs have RNA binding specificities that overlap but are also distinct from each other. This suggests that the expression levels of these factors in different tissues may play an important role in alternative polyadenylation. Furthermore, CstF64 and CstF64τ are likely to have related roles in poly(A) site selection, as depleting both factors leads to a greater number of changes in alternative polyadenylation compared to depleting CstF64 alone (Yao et al. 2013).

CstF77 forms a homodimer, and it serves as a scaffold protein within the CstF complex. As CstF50 and CstF64 do not directly interact with each other, CstF77 bridges this interaction through its proline rich region (Takagaki and Manley 2000). Recently, CstF77 has also been shown to increase the binding affinity of CstF64 to RNA (Yang et al. 2018).

Lastly, CstF50 also forms a homodimer and it interacts with CstF77. CstF50 can recognize U/GU-rich sequences with different length and content and by doing so, affecting the specificity of CstF-RNA interaction (Yang et al. 2018). CstF50 also interacts with the C-terminal domain (CTD) of RNA polymerase II (RNAPII). This CstF-RNAPII interaction is important for the coupling of transcription termination and mRNA 3’ processing (McCracken et al. 1997; Mandel et al. 2007).

CFIm:

The mammalian cleavage factor I (CFIm) is a tetrameric complex consisting of a homodimer of the small subunit, CFIm25, and a homodimer of one of two alternative larger subunits, either CFIm59 or CFIm68 (Rüegsegger et al. 1998). Recently it has been shown that the CFIm complex is not required for either cleavage or polyadenylation. (Mandel et al. 2007; Chan et al. 2011; Boreikaite et al. 2022). Rather, it serves as a sequence-dependent activator for UGUA-containing poly(A) sites (usually enriched at distal poly(A) sites) and is essential for alternative polyadenylation (Zhu et al. 2018). CFIm is essential for both RNA binding and recruitment of other 3’ processing factors. The RNA-protein and protein-protein interactions of CFIm will be further discussed in this section.

CFIm in RNA Binding:

All three subunits of CFIm have been shown to interact with RNA through UV-crosslinking experiments, indicating their involvement in RNA recognition (Rüegsegger et al.

1996). However, the small subunit CFIm25 is known to directly bind to the UGUA enhancer motif, which is typically located about 50 nucleotides upstream of distal poly(A) sites. CFIm25 belongs to the Nudix hydrolase superfamily of proteins and has a Nudix domain with the classic α/β/α fold. However, CFIm25 is unique among other Nudix hydrolase proteins in that it lacks two critical glutamate residues, rendering it catalytically inactive (Coseno et al. 2008). Instead, the Nudix domain of CFIm25 directs its binding to the UGUA upstream enhancer sequences (Yang et al. 2010).

While the large subunits of CFIm, CFIm59 and CFIm68 also have the ability to directly crosslink to RNA, the mechanism for RNA recognition is not yet fully understood (Rüegsegger et al. 1996; Martin et al. 2012). Although both subunits contain an RNA recognition motif, this motif is actually important for interaction with CFIm25 rather than for direct RNA binding. Therefore, it is likely that CFIm59 and CFIm68 play a role in regulating the RNA binding specificity of CFIm25. Further investigation is needed to determine the RNA binding specificity of CFIm59 and CFIm68.

CFIm in recruitment of other 3' processing factors:

Numerous studies have demonstrated that CFIm contributes to alternative polyadenylation. CFIm knockdown in cells induces widespread changes in poly(A) site usage, resulting in a shift from distal (downstream) to proximal (upstream) poly(A) sites (Martin et al. 2012; Masamha et al. 2014; Zhu et al. 2018). Our lab recently reported that CFIm stimulates APA by activating distal poly(A) sites (Zhu et al. 2018). When CFIm binds to the UGUA enhancer sequence located upstream of distal poly(A) sites, it recruits CPSF to the downstream AAUAAA hexamer through interactions between the RS domain of CFIm59 or CFIm68 and the

RE/D domain of Fip1, a subunit in the CPSF complex. As a result, when CFIm levels are high in the cell, there is a preference for the usage of the distal, UGUA-containing poly(A) site (Zhu et al. 2018).

CFIIm:

The mammalian cleavage factor II complex (CFIIm) consists of two proteins: Pcf11 and Clp1. CFIIm is a core member of the mRNA 3′ processing complex but was not originally thought to be a stably associated complex based on immunoprecipitation and RNA pull down experiments. It was recently shown to be essential for the activation of pre-mRNA cleavage in vitro using recombinant proteins (Boreikaite et al. 2022). Pcf11 interacts with the RNA polymerase II C-terminal domain (RNAPII CTD) in a phosphorylation-dependent manner through its CTD-interacting domain (CID). Mutation of Pcf11 CID blocks transcription termination but not pre-mRNA cleavage, suggesting a role of Pcf11 in regulating RNAPII transcription termination (Sadowski et al. 2003; Proudfoot 2004; Kamieniarz-Gdula et al. 2019). Pcf11 can be autoregulated by a proximal, intronic poly(A) site that inhibits the formation of full-length Pcf11 transcripts (Wang et al. 2019). In yeast and Drosophila, Pcf11 homologs are essential for both pre-mRNA 3′ processing and transcription termination (Proudfoot 2004). Furthermore, the yeast homolog Pcf11p connects pre-mRNA 3′ processing to mRNA export by recruiting the mRNA export adaptor protein Yra1p to the mRNA (Johnson et al. 2009).

The role of Clp1 in mRNA 3′ processing is not well understood. Clp1 amino acids R288 and R293 were identified as ATP binding sites in the protein. However, the ATP binding ability and Clp1-Pcf11 interaction are two separate properties since recombinant Clp1 mutants that cannot bind to ATP can still be purified as a stable complex together with Pcf11 (Schmidt et al. 2022).

Rbbp6:

Rbbp6 was first identified as a co-purifying protein when the mRNA 3′ processing complex was characterized using an RNA pull down approach. However, unlike it's yeast homolog MpeI, which binds to yeast CPSF73 homolog Ysh1 and stably associates with the yeast cleavage and polyadenylation complex, the human Rbbp6 is thought to be a non-stably associated factor (Shi et al. 2009). Two recent studies suggested that Rbbp6 is an essential factor for pre-mRNA cleavage using in vitro assay with recombinantly purified proteins (Boreikaite et al. 2022; Schmidt et al. 2022). A cryo-EM structural study on Rbbp6-CPSF complex suggested that Rbbp6 directly binds to CPSF73, and its N-terminal 335 amino acids are sufficient for this interaction (Schmidt et al. 2022). As CPSF73 normally adopts a closed conformation, the interaction between Rbbp6 and CPSF73 is believed to be an essential step during mRNA 3′ processing to open up CPSF73, thereby preparing it for the cleavage of pre-mRNA (Hill et al. 2019).

PAP:

Poly(A) polymerase (PAP) is important for the addition of a poly(A) tail following pre-mRNA cleavage step in 3′ processing. By itself, PAP binds to RNA through its RNA binding domain and polyadenylates RNA in a nonspecific and AAUAAA-independent manner. The specificity of PAP is achieved by its interaction with other 3' processing factors, specifically, through the interaction with Fip1 in the CPSF complex (Chan et al. 2011). The length of poly(A) tail synthesized by PAP varies in different genes and under different condition. In vitro, PAP synthesis a poly(A) tail around 200 nt. The poly(A) tail length in cells range from ~30 nt to ~250 nt, with shorter poly(A) tails usually found in highly expressed genes (Lima et al. 2017).

1.2 MECHANISM AND REGULATION OF ALTERNATIVE POLYADENYLATION

mRNA 3′ processing is a tightly regulated process in the gene expression pathway. In humans, it is estimated that more than 70% of genes contain multiple poly(A) sites. As a result, these genes can be cleaved and polyadenylated at multiple alternative poly(A) sites. This process is called alternative polyadenylation (APA). There are two types of APA: 3′ UTR APA and intronic APA (IPA) (Figure 1.3). In this section I will discuss these two types of APA in detail, including their regulatory mechanisms and biological consequences.

3′ UTR APA:

3′ UTR APA takes place when all of the poly(A) site lies within the 3′ UTR region (also refered to as the terminal exon in some cases) in the pre-mRNA. In general, the distal (downstream) poly(A) sites are usually stronger in comparison to the proximal (upstream) poly(A) site. This is because of the enrichment of cis elements, including the UGUA enhancer sequences and/or U/GU-rich downstream sequences. Although the proximal poly(A) sites are usually weaker, they are transcribed first, so this may serve as an inherent advantage for the proximal poly(A) sites to allow assembly of the 3′ processing complex while the distal poly(A) sites are still being transcribed (Davis and Shi 2014).

3′ UTR APA generates RNA transcripts that contain the same protein coding region but different length 3′ UTRs, which are hotspots for RNA binding proteins and microRNAs. When different RNA binding proteins and/or microRNAs are bound to the 3′ UTR, the RNA transcript can be regulated differently, resulting in difference in RNA stability, translation efficiency, and/or mRNA/protein localization (Tian and Manley 2016).

Intronic APA:

Another form of APA is known as intronic APA, which occurs when a polyadenylation site is located within the intronic region of a gene. Polyadenylation at these intronic sites often results in the production of truncated proteins, which may occasionally be functional, but are more commonly subjected to degradation. In instances where these truncated proteins are functional, intronic APA contributes to an increase in protein diversity and can yield protein isoforms with distinct functions (Soles and Shi 2021).

A notable example of intronic APA's role in generating functionally different protein isoforms can be observed during B cell differentiation. In this process, usage of the distal poly(A) site located within the 3′ UTR of the immunoglobulin M (IgM) heavy chain gene gives rise to membrane-bound antibodies (Takagaki et al. 1996). In contrast, when the proximal intronic poly(A) site is used, secreted antibodies are generated. This highlights the importance of intronic APA in modulating protein functionality and diversifying the cellular proteome.

Regulation of APA:

The abundance of mRNA 3′ processing factors within a cell plays a crucial role in the regulation of APA. As mentioned in an earlier section, the CFIm complex interacts with UGUA enhancer sequences, which are typically enriched at distal poly(A) sites. By binding to these UGUA-containing sequences, CFIm promotes APA to occur at distal poly(A) sites (Zhu et al. 2018). In addition to CFIm, another 3' processing factor, CstF64, also plays a role in the regulation of APA. During the process of B cell differentiation, the choice between proximal intronic poly(A) sites µS or distal 3′ UTR poly(A) sites µM is influenced by the level of CstF64 present in the cell. In B cells, CstF64 levels are relatively low, which leads to the preferential use of distal poly(A) sites µM. Conversely, in plasma cells where CstF64 levels are higher, the proximal poly(A) site µS is favored (Takagaki et al. 1996). This dynamic interplay between the levels of mRNA 3' processing factors and the choice of polyadenylation sites highlights the intricate regulatory mechanisms that govern APA. By influencing the selection of poly(A) sites, these factors contribute to the generation of diverse mRNA isoforms and, consequently, a variety of protein isoforms with distinct functions.

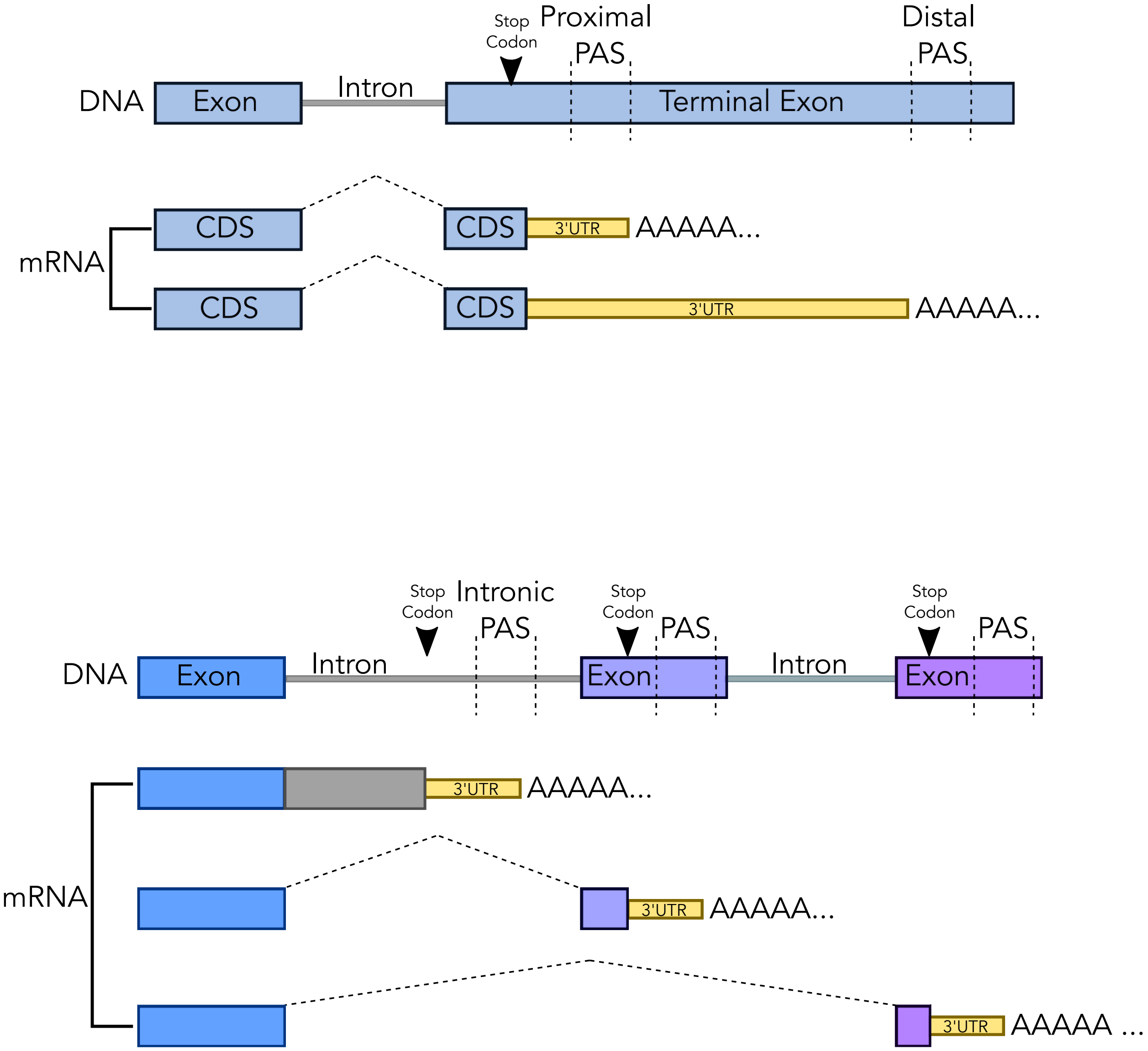

Figure 1.3 Alternative Polyadenylation (APA)

Schematic representation of two types of alternative polyadenylation: 3′ UTR APA (top) and intronic APA (bottom)

1.3 MODULATION OF mRNA 3′ PROCESSING DURING VIRAL INFECTION

As a crucial stage in the gene expression pathway, mRNA 3′ processing is subjected to stringent regulation. A myriad of viruses employ diverse strategies to interfere with the host's mRNA 3′ processing, with the ultimate goal of either shutting down host gene expression to facilitate the expression of their own viral genes or commandeering the host machinery to aid in viral 3′ processing (Vijayakumar et al. 2022). Given that mRNA 3′ processing is coupled to transcription termination by RNAPII, any disruption to mRNA 3′ processing can lead to aberrant transcription termination. In such cases, RNAPII continues to transcribe beyond the transcription end site of genes, resulting in the generation of long readthrough transcripts. In this section, I will provide a summary of six viruses that are currently known to interfere with mRNA 3′ processing and/or transcription termination. This information will serve as a foundation for understanding the various tactics employed by different pathogens and ultimately contribute to the development of novel therapeutic strategies to combat various viral infection.

Influenza A Virus (IAV):

Influenza A virus (IAV) is an enveloped, negative-sense, single-stranded RNA (ssRNA) virus belongs to the *Orthomyxoviridae* that causes seasonal epidemics of flu disease in people. The viral nonstructural protein NS1 of IAV has been shown by multiple groups to modulate mRNA 3′ processing and transcription termination. In vitro, purified recombinant GST-NS1 protein prevents mRNA 3′ processing complex formation and inhibits both cleavage and polyadenylation of pre-mRNA substrates. Likewise, in cells, NS1 protein also inhibited both cleavage and polyadenylation of reporter poly(A) sites (Nemeroff et al. 1998). The NS1 amino acids 144 and 184-188 interact with zinc finger 2 and 3 of CPSF30, a core subunit of the CPSF complex (Twu et al. 2006). Interestingly, the zinc finger 2 and 3 of CPSF30 are also required for the direct interaction with AAUAAA hexamer within the poly(A) site. Zinc finger 2 binding to A1, A2 and zinc finger 3 binding to A4, A5 (Chan et al. 2014; Sun et al. 2018). This suggests that influenza virus may have evolved a way to specifically target CPSF30-AAUAAA interaction to suppress host mRNA 3′ processing.

In addition to targeting CPSF30 as a way to shutoff host mRNA 3′ processing, IAV infection also induces increased RNAPII occupancy downstream of the transcription end site within protein coding genes and transcription termination defect in cells. This leads to the downregulation of various classes of protein-coding genes that play critical roles in cell division, apoptosis, cell defense, and metabolism (Zhao et al. 2018; Bauer et al. 2018).

<u>Herpes Simplex Virus Type 1 (HSV-1):</u>

Herpes Simplex Virus Type 1 (HSV-1) is an enveloped, double-stranded DNA (dsDNA) virus in the *Herpesviridae* family that infects diverse metazoans. The symptoms of HSV-1 infection range from painful skin lesions or cold sores around the mouth to life threatening encephalitis (Whitley and Roizman 2001). Previously our lab and others reported that HSV-1 infection leads to widespread transcription termination defect in cells, and the viral immediate early protein ICP27 plays a key role in HSV-1-induced disruption of transcription termination (DoTT) (Rutkowski et al. 2015; Wang et al. 2020). Mechanistically, HSV-1 ICP27 protein has a bimodal activity. On the one hand, ICP27 interacts with the CPSF complex during mRNA 3′ processing through its C-terminal domain. This interaction displaces Symplekin from the 3′

processing complex, leading to formation of a defective 3′ processing complex on host genes, thereby inhibiting cleavage of host pre-mRNA and causing a transcription termination defect in virus infected cells. On the other hand, ICP27 also binds to poly(A) site containing GC-rich upstream sequences, which are commonly found in the viral genome (68.3% GC) compared with the host genome (40.1% GC). When ICP27 binds to the upstream GC-rich sequences, it becomes a sequence-dependent activator and activates 3' processing of that specific poly(A) site (Wang et al. 2020).

It is worth noting that ICP27 may not be the only HSV-1 protein that can induce a transcription termination defect in virus infected cells. ICP4, another HSV-1 immediate early protein, may also contribute to HSV-1-induced DoTT as infection with a ICP4 deleted virus (HSV-1 ΔICP4) greatly reduced transcription readthrough (Wang et al. 2020). More studies need to be performed to understand the exact role of ICP4 on mRNA 3′ processing and transcription termination.

Human Cytomegalovirus (HCMV):

Human Cytomegalovirus (HCMV) is a beta-herpesvirus that also belongs to the *Herpesviridae* family. Symptoms of HCMV infection are usually mild or asymptomatic in healthy individuals but more sever in babies or immunocompromised people, ranging from pneumonia, hearing loss, to encephalitis. HCMV infection induces widespread APA changes. Specifically, these APA changes resulted in 3′ UTR shortening. The host protein CPEB1 is up regulated during infection, and siRNA knockdown of CPEB1 lengthened the 3′ UTR, leading to partial or full rescue in ~40% of genes that showed APA changes (Batra et al. 2016).

Vesicular Stomatitis Virus (VSV):

Vesicular Stomatitis Virus (VSV) is an enveloped, ssRNA virus in the *Rhabdoviridae* family. VSV primarily infects livestock such as horses and cattles, although humans can also be infected through contact with infected animals. VSV infection in livestock leads to the disease vesicular stomatitis whereas infection in human leads to mild flu like symptoms (Ludwig and Hengel 2009). In both human and mouse cells infected with VSV, an average 3′ UTR shortening is observed for host mRNAs. Gene Ontology (GO) analysis revealed that the genes displaying APA are enriched in immune-related categories, highlighting potential role of mRNA 3' processing in the antiviral immune response (Jia et al. 2017).

Enterovirus A71 (EV-A71):

Enterovirus A71 (EV-A71) is a positive-sense, ssRNA virus that belongs to the family of *Picornaviridae*. It is one of the major causative agents for hand, foot, and mouth disease (HFMD) and is also associated with severe neurological manifestations such as acute flaccid myelitis (Nguyen-Tran and Messacar 2022). The viral 3C^{pro} protease cleaves CstF64 within the mRNA 3′ processing complex in vitro at amino acid 251 and multiple positions between amino acid 483-515. Viral 3C^{pro} treated HeLa nuclear extract (NE) also showed defects in both in vitro cleavage and in vitro polyadenylation. During EV-A71 infection, CstF64 levels are reduced in virus infected cells and there is an accumulation of a reporter pre-mRNA and reduction of the reporter polyadenylated RNA (Nguyen-Tran and Messacar 2022). However, as this previous study only used a reporter pre-mRNA, the genome wide effect of EV-A71 infection on mRNA 3′ processing and transcription termination remains to be characterized.

Human Immunodeficiency Virus-1 (HIV-1):

Human immunodeficiency virus-1 (HIV-1) is an enveloped, positive-sense, ssRNA virus belongs to the *Retroviridae* family. HIV-1 is the causative agent for AIDS, which remains a leading global health concern. Overexpression of HIV-1 Tat protein, the trans activator of transcription, leads to up regulation of CPSF73, both on the RNA and protein level (Calzado et al. 2004). Co-immunoprecipitation of Tat in Tat overexpression cells indicated that Tat interacts with CPSF73, the endonuclease within the 3′ processing complex (de la Vega et al. 2007). The effect of Tat on mRNA 3′ processing and transcription termination in HIV-1 infected cells remains an unexplored area. However, as Tat interacts with and increases the expression level of CPSF73, it is possible that the HIV-1 Tat protein recruits CPSF73 to sites of viral mRNA 3′ processing, thereby facilitating the expression of HIV genes. More research is required to test this possibility.

Besides CPSF73, HIV-1 also interacts with CFIm68, which is thought to be important for integration of HIV genome into the host genome (Sowd et al. 2016; Bejarano et al. 2019; Zheng et al. 2021). HIV-1 can also induce the formation of CFIm68 biomolecular condensates (Ay and Di Nunzio 2023). However, how these processes affect the host mRNA 3′ processing and transcription termination is unclear.

At present, these are the viruses that have been identified as having the ability to modulate host mRNA 3′ processing and/or transcription termination. However, the fact that these viruses belong to diverse viral families highlights that the capability to interfere with mRNA 3’ processing is not restricted to a specific family or type of virus. This also suggests that viral manipulation of host mRNA 3’ processing could be a widespread strategy employed by various viruses to enhance their replication and spread within the host. Future research focusing on virus-mediated regulation of mRNA 3′ processing is highly likely to uncover additional viruses that are capable of manipulating this critical step in the gene expression pathway. These investigations will not only improve our understanding of viral pathogenesis but also aid in the development of novel antiviral therapies.

1.4 mRNA 3′ PROCESSING AS AN ATTRACTIVE DRUG TARGET

The recent discovery and characterization of small molecule inhibitors of mRNA 3′ processing has revealed a promising strategy for targeting mRNA 3′ processing as a treatment for protozoan parasite infections and specific cancers. There are currently two small molecule inhibitors: AN3661 and JTE-607, both of which target the endonuclease CPSF73 within the 3′ processing complex. In this section I will provide the current knowledge on these inhibitors.

AN3661:

AN3661 is a benzoxaborole compound that specifically targets the protozoan but not human CPSF73. In vitro, AN3661 potently inhibits the growth of multiple protozoan parasites including *Toxoplasma*, *Plasmodium*, *Trypanosoma*, and *Cryptosporidium* (Sonoiki et al. 2017; Palencia et al. 2017; Wall et al. 2018; Swale et al. 2019). An in vivo study using mice infected with *Toxoplasma gondii* indicated that AN3661 treatment eliminated the infection and no signs of illness were found in the mice (Palencia et al. 2017). A structural study revealed that AN3661 binds to the parasite CPSF73 within the active site. The specific inhibition of parasite but not human CPSF73 is possibly achieved by the slight structural differences between the two homologs, with the parasite CPSF73 containing a long loop along the metallo-β-lactamase domain but the human CPSF73 does not (Swale et al. 2019).

JTE-607:

JTE-607 is a human mRNA 3′ processing inhibitor. It was discovered over two decades ago as a multiple cytokine production inhibitor and has been shown effective in cell culture, a mouse model, and healthy human volunteers (Kakutani et al. 1999; Jian et al. 2004; Ryugo et al. 2004; Borozdenkova et al. 2011). Additionally, JTE-607 is also an anti-cancer drug as it prolonged the survival in a mouse acute myeloid leukemia (AML) model, despite the lack of knowledge on its intracellular target (Uesato et al. 2006). Recently, using an unbiased screen coupled with chemical genetics, the intracellular target of JTE-607 was identified to be CPSF73 (Ross et al. 2020; Kakegawa et al. 2019). JTE-607 binds to human CPSF73 within its active site. Binding of JTE-607 and CPSF73 blocks pre-mRNA cleavage during 3′ processing, leading to transcription termination defect and DNA-RNA R-loop formation (Ross et al. 2020). Since the discovery of CPSF73 as the intracellular target of JTE-607, several studies have revealed that various cancers exhibit sensitivity to JTE-607. These include acute myeloid leukemia (AML), Ewing sarcoma, pancreatic ductal adenocarcinoma, and triple-negative breast cancer (Uesato et al. 2006; Ross et al. 2020; Alahmari et al. 2022; Liu et al. 2022).

The recent identification of small molecule inhibitors AN3661 and JTE-607 has revealed a promising strategy for targeting mRNA 3′ processing in the treatment of protozoan parasite infections and select cancers. This innovative approach represents a previously unexplored avenue for the development of new therapeutic interventions.

1.5 SUMMARY

mRNA 3′ processing is a crucial and intricate step in the gene expression pathway that is subject to stringent regulation. This section outlined our current understanding of the mechanisms and regulatory processes involved in mRNA 3′ processing and alternative polyadenylation. Additionally, the impact of viral infections and small molecule inhibitors on mRNA 3' processing has been described. However, numerous questions remain unanswered. The following work is dedicated to dissecting the molecular mechanisms that govern mRNA 3′ processing regulation during HSV-1 virus infection and in the presence of the small molecule inhibitor JTE-607.

Chapter 2 built upon our prior research on HSV-CPSF interactions and employed various high-throughput sequencing techniques to investigate the mechanisms and outcomes of APA regulation during HSV-1 infection. We have characterized the types of APA changes following viral infection and examined the fate of the distinct APA isoforms induced by HSV-1 infection. Chapter 3 utilized multiple complementary approaches, including in vitro biochemistry, machine learning, and high-throughput sequencing, to examine the mechanisms underlying JTE-607-mediated mRNA 3′ processing inhibition. We have developed a machine learning-based tool for accurately predicting poly(A) site sensitivity to JTE-607 and have proposed a model explaining the sequence-specific inhibition of mRNA 3' processing by the small molecule inhibitor JTE-607.

CHAPTER 2

Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation

2.1 Publication Note:

This paper was originally published as an open access article in PLoS Genetics under the terms of the Creative Commons Attribution License. Liang Liu, the author of this dissertation, is the copyright owner and is the co-first author of this publication.

Wang, X.*, **Liu, L.***, Whisnant, A.W., Hennig, T., Djakovic, L., Haque, N., Bach, C., Sandri-Goldin, R.M., Erhard, F., Friedel, C.C., Dölken, L., Shi, Y. (2021) Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation. PLoS Genet 17(3): e1009263.

2.2 Summary

Eukaryotic gene expression is extensively regulated by cellular stress and pathogen infections. We have previously shown that herpes simplex virus 1 (HSV-1) and several cellular stresses cause widespread disruption of transcription termination (DoTT) of RNA polymerase II (RNAPII) in host genes and that the viral immediate early factor ICP27 plays an important role in HSV-1-induced DoTT. Here, we show that HSV-1 infection also leads to widespread changes in alternative polyadenylation (APA) of host mRNAs. In the majority of cases, polyadenylation shifts to upstream poly(A) sites (PAS), including many intronic PAS. Mechanistically, ICP27 contributes to HSV-1-mediated APA regulation. HSV-1- and ICP27- induced activation of intronic PAS is sequence-dependent and does not involve general inhibition of U1 snRNP. HSV1-induced intronic polyadenylation is accompanied by early termination of RNAPII. HSV-1-induced mRNAs polyadenylated at intronic PAS (IPA) are exported into the cytoplasm while APA isoforms with extended 3’ UTRs are sequestered in the nuclei, both preventing the expression of the full-length gene products. Finally, we provide evidence that HSV-induced IPA isoforms are translated. Together with other recent studies, our results suggest that viral infection and cellular stresses induce a multi-faceted host response that includes DoTT and changes in APA profiles.

2.3 Introduction

The 3′ ends of the vast majority of eukaryotic mRNAs are formed through cleavage and polyadenylation (Colgan and Manley 1997; Zhao et al. 1999; Chan et al. 2011). In mammals, poly(A) sites (PAS) are defined by several cis-elements, including the AAUAAA hexamer, the U/GU-rich downstream element, and other auxiliary sequences. These sequences recruit RNA 3′ processing factors CPSF, CstF, CFIm, CFIIm, and the poly(A) polymerase to form the 3′ processing complex. RNA 3′ processing occurs cotranscriptionally and it plays an essential role not only in RNA biogenesis, but also in transcription termination by RNA polymerase II (RNAPII) (Richard and Manley 2009; Eaton et al. 2020; Proudfoot 2016). According to the "allosteric model" of transcription termination, the transcription machinery undergoes a transformation upon passing through a PAS, which primes RNAPII for termination. Alternatively, the "torpedo model" posits that the unprotected 5' end of RNA generated by the 3′ processing cleavage step is recognized by the exoribonuclease Xrn2. Xrn2-mediated degradation of the nascent RNA ultimately leads to transcription termination. Thus, in both models, RNA 3′ processing plays a central role in transcription termination.

RNA 3′ processing also plays an important role in gene regulation. The transcripts of over 70% of human genes can be cleaved and polyadenylated at multiple alternative PAS, a process called alternative polyadenylation (APA) (Tian and Manley 2016; Shi 2012; Mayr 2016). Different APA isoforms from the same gene may encode distinct proteins and/or contain different 3' untranslated regions (UTRs). 3′ UTRs are hot spots for regulation: they harbor target sites for microRNAs, binding sites for RNA- binding proteins, RNA localization signals, and they can function as protein assembly platforms. Thus, APA isoforms from the same gene could be differentially regulated. Recent studies have provided evidence that APA plays important roles in a wide variety of biological processes and aberrant APA regulation has been linked to a number of diseases, including cancer and neurological disorders (Di Giammartino et al. 2011). Many APA regulators have been identified, including the core RNA 3′ processing factors, splicing factors, and RNA-binding proteins (Shi and Manley 2015). For example, U1 snRNP has been shown to inhibit premature cleavage/polyadenylation at intronic PAS, thereby protecting transcript integrity globally (Kaida et al. 2010). Despite recent progress, however, the regulatory mechanisms and functional consequences of APA remain poorly understood.

Both RNA 3′ processing and transcription termination are highly regulated. For example, we have previously shown that HSV-1 infection leads to a widespread disruption of transcription termination (DoTT) (Wang et al. 2020; Rutkowski et al. 2015). Influenza virus (IAV) was reported to elicit a similar response (Zhao et al. 2018). The Steitz lab observed a transcription termination defect in cells exposed to salt/osmotic stress that leads to the production of transcripts downstream of genes (DoGs) (Vilborg et al. 2015). A comparative analysis showed that virus-induced DoTT and stress-induced DoGs are highly related (Hennig et al. 2018). Although the mechanism for DoTT/DoGs remains unclear, we have recently shown that the viral immediate early factor ICP27 contributes to HSV-1-induced DoTT by directly binding to the RNA 3′ processing factor CPSF and inhibiting the cleavage step (Wang et al. 2020). Meanwhile, several groups reported that virus infections, such as the human cytomegalovirus (HCMV) and vesicular stomatitis virus (VSV), or stress can induce global APA changes (Batra et al. 2016; Jia et al. 2017; Zheng et al. 2018). However, the relationship between virus- or stress-induced APA and DoTT/DoG remains unclear. In this study, we integrated time-resolved global APA profiling, nascent RNA sequencing, cell fractionation and RNA sequencing, and ribosome profiling (Ribo-seq) data in HSV-1-infected cells to elucidate the scope, mechanism, and functional impact of virus-induced APA changes and DoTT.

2.4 Results

HSV-1 infection induces widespread and dynamic APA changes

To determine if and how the global APA profile of host genes is altered during HSV-1 infection, we performed PAS-seq analysis of HeLa cells at 0, 2, 6, and 12 hours post-infection (hpi). PAS-seq is a method developed in our laboratory for quantitatively mapping RNA poly(A) junctions (Shepard et al. 2011). Briefly, poly(A)+ RNAs are fragmented to ~200 nucleotide (nt) fragments and reverse transcribed using oligo(dT) primers, and then the poly(A) junction-containing DNAs are amplified for high throughput sequencing. This method has been used extensively for pro- filing global APA (Shepard et al. 2011; Zhu et al. 2018). By comparing the APA profiles of cells at 0 and 12 hpi, we detected significant APA changes in 1,050 genes (FDR < 0.05, impacting at least 15% of transcripts, see Methods for details). In 745 genes (71%), polyadenylation shifted to proximal PAS (Distal-to-Proximal or DtoP) and 305 (29%) showed changes in the opposite direction (Proximal-to-Distal or PtoD) (Figure 2.1A). HSV-1 infection led to the apparent activation of many PAS that are unused in uninfected cells. For example, among the DtoP changes, 188 genes (25%) shifted to a proximal PAS that was not used in uninfected cells. 44 genes (14%) of those displayed PtoD shifts activated a previously unused distal PAS. Additionally, a significant portion of HSV-1-induced changes involved intronic PAS. For example, among the DtoP changes, 331 (44%) shifted to a proximal intronic PAS

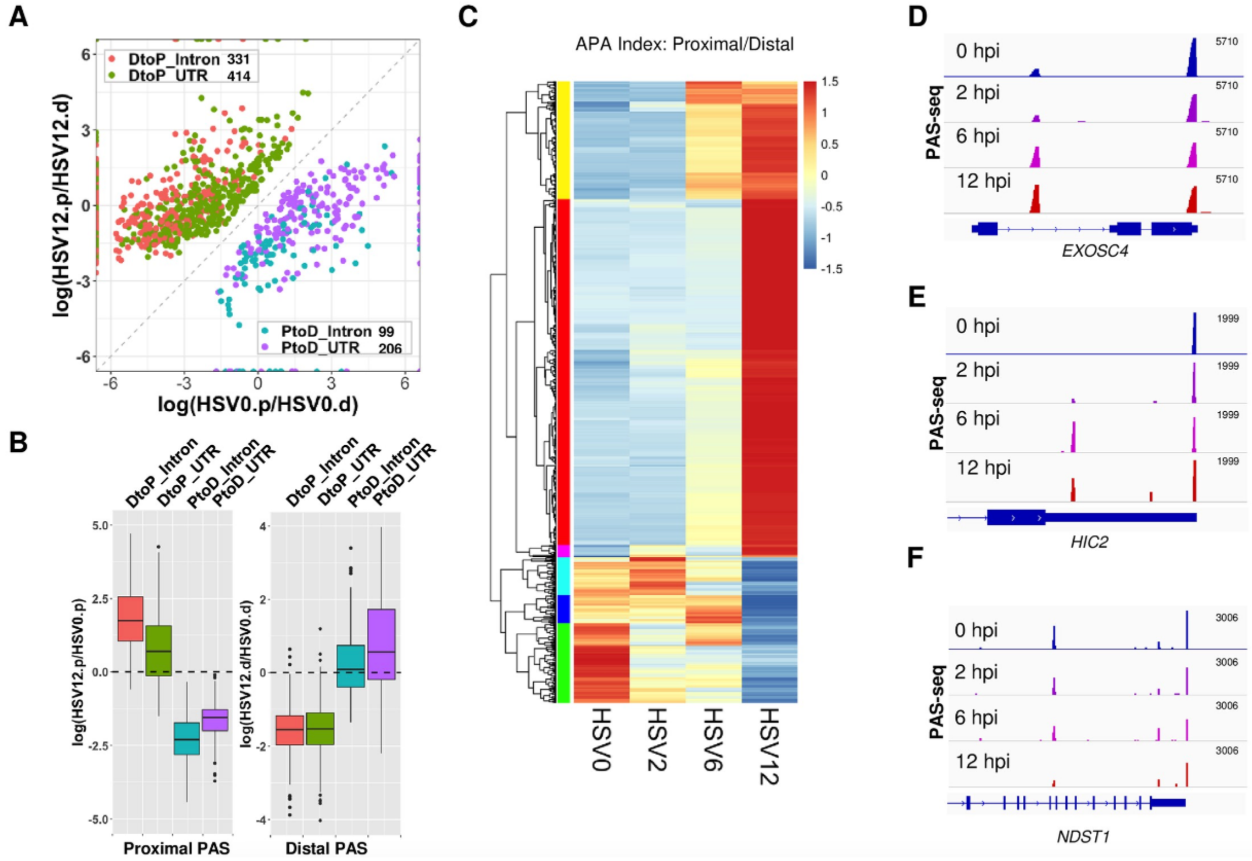

Figure 2.1 HSV-1 infection induces widespread and dynamic APA changes

(A) A scatter plot of HSV-1-induced significant APA changes in HeLa cells (FDR < 0.05, at least 15% of the transcripts shifted). HSV0.p or HSV12.p: read counts for the proximal PAS in cells at 0 or 12 hpi; HSV0.d or HSV12.d: read counts for the distal PAS in cells at 0 or 12 hours post infection. DtoP: a distal to proximal shift; PtoD: a proximal to distal shift. Intron: shifts involving an intronic PAS. UTR: both PAS are located in the 3′ UTR.
(B) Relative read count changes at proximal (.p) or distal (.d) PAS in the four groups of APA shifts.
(C) A heat map showing the APA index (proximal/distal read count ratio) of all the genes with significant APA shifts as shown in (A). Data was scaled by row. Color bars on the left denote the 6 groups that displayed similar kinetic patterns.
(D-F) PAS-seq tracks of example genes.

(DtoP_Intron) while 32% of PtoD genes shifted from a proximal intronic PAS to a PAS in the 3′ UTR (PtoD_Intron, Figure 2.1A). APA profile changes could be due to differential PAS selection during transcription, differential degradation of APA isoforms, a selective loss of proximal or distal PAS due to read-through transcription, or a combination of all factors. To begin to understand the cause of the observed APA changes in HSV-1-infected cells, we compared the PAS-seq reads (normalized by sequencing depths of the host transcriptomes) at proximal and distal PAS of genes that displayed significant APA changes. As shown in Figure 2.1B, DtoP changes were accompanied by a relative increase in proximal PAS reads and a relative decrease in distal PAS reads. Conversely, PtoD changes involving only UTR PAS were accompanied by the opposite changes. In addition to ratios, DtoP changes were accompanied by a net increase in the normalized read counts at proximal PAS and a net decrease at the distal PAS, while the opposite changes were observed for PtoD shifts (Figure 2.2). These results provided evidence that the HSV-1-induced APA changes are not caused solely by preferential degradation or loss of specific APA isoforms, but may require a shift in PAS usage. The underlying mechanism will be further addressed below.

To determine the kinetics of APA changes during HSV-1 infection, we compared the APA index (read count ratio between proximal and distal PAS) of the 1,050 genes. The greatest shifts in APA profile occurred between 6 and 12 hpi (Figure 2.1C). However, multiple different kinetic patterns were observed for the timing and magnitude of APA changes (Fig 1C, see the colored sidebars for classification), indicating that multiple mechanisms are involved in regulating the APA of host genes. Three examples were provided to illustrate the different kinetic groups. For example, polyadenylation of the EXOSC4 transcripts shifted from a PAS in the 3′ UTR to a proximal intronic PAS (Figure 2.1D). The majority of the APA shift occurred

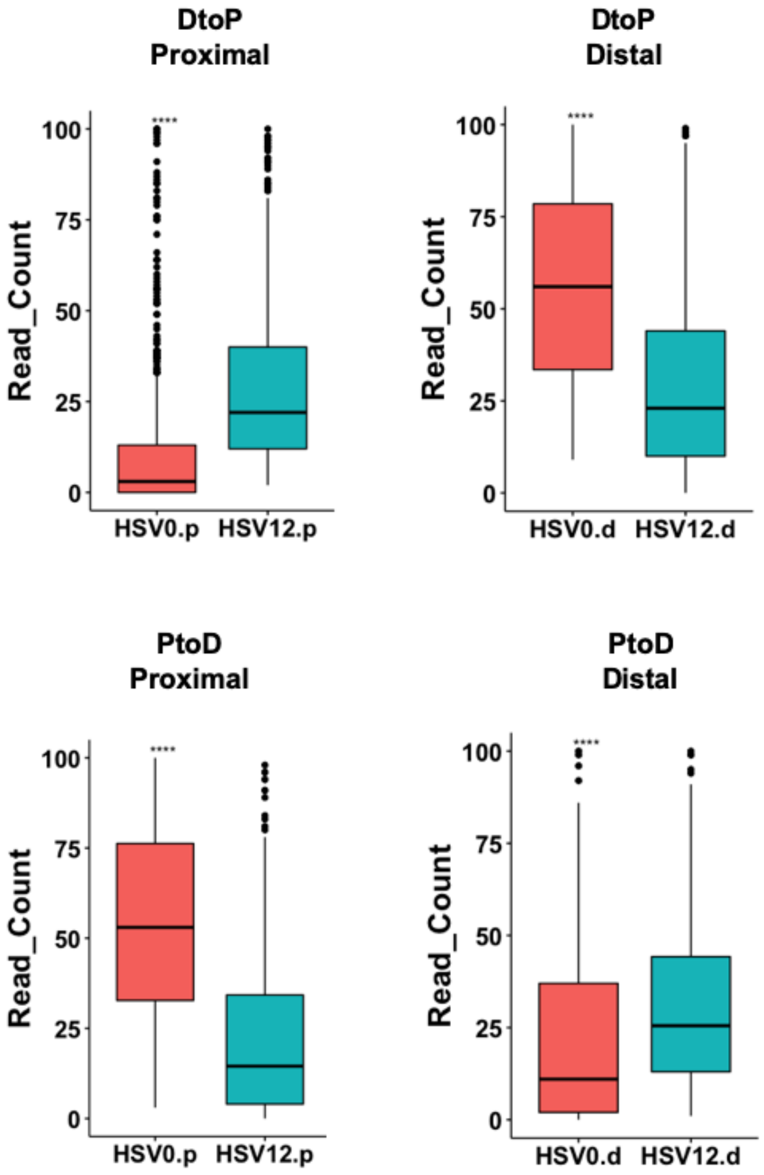

Figure 2.2 Normalized PAS-seq read counts at proximal and distal PAS of APA genes

between 2 and 6 hpi and a modest further shift was observed between 6 and 12 hpi. Similarly, polyadenylation of HIC2 transcripts shifted to a proximal intronic PAS between 2 and 6 hpi. However, the usage of this intronic PAS decreased subsequently (Figure 2.1E). Finally, a PtoD shift was observed for NDST1 transcripts and the majority of the APA change occurred between 6 and 12 hpi (Figure 2.1F). Together, these data demonstrated that HSV-1 infection induces widespread APA changes, the majority of which shift from distal to proximal PAS. These APA changes follow multiple kinetic patterns, indicating that different mechanisms might be involved in HSV-1-mediated APA regulation of host genes.

The relationship between the HSV-1-induced APA changes and transcription

HSV-1-induced APA changes could be due to changes in PAS selection during transcription and/or selective loss of individual APA isoforms. To distinguish between these mechanistic models, we directly compared our PAS-seq data with nascent RNA sequencing (4sU-seq) data, which provides information on transcription activities (Wang et al. 2020). We focused our analyses on the gene body as well as 1 kilobase (kb) downstream of the transcript end site (TES) in order to monitor both transcription elongation and termination. To avoid detecting signals from neighboring genes, we selected the APA genes that do not overlap with other genes within the 1 kb downstream region (508 DtoP and 130 PtoD genes). Meta-analyses of 4sU-seq signals in mock or HSV-1-infected cells along the genes that showed HSV-1-induced higher usage of upstream PAS (DtoP) revealed two interesting differences. First, although the 4sU-seq signals were similar at transcription start sites (TSS), the signal intensities were significantly lower within the gene body in HSV-1-infected cells (Figure 2.3A, p values for each position were calculated using Wilcoxon rank sum method and shown as a color-coded bar

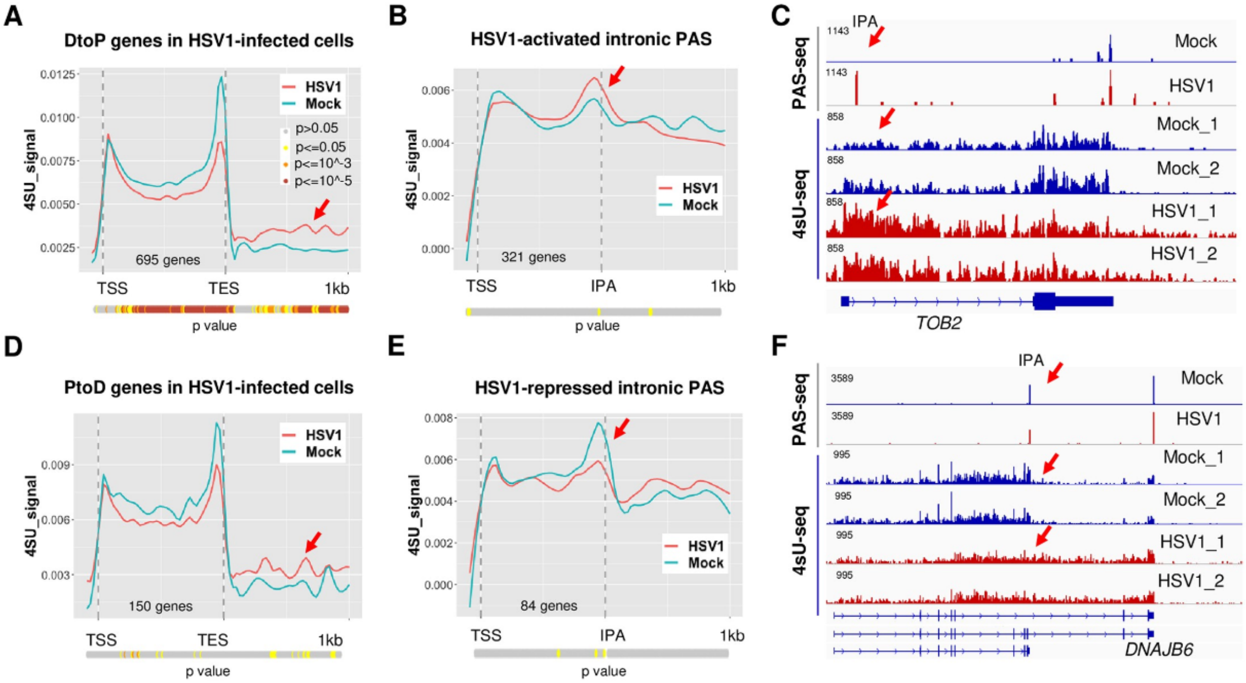

Figure 2.3 HSV-1-induced APA changes occur, at least in part, co-transcriptionally

(A) 4sU-seq signals of genes that displayed DtoP APA shifts in HSV-1-infected HeLa cells (8 hpi). TSS: transcription start site. TES: transcription end site. P values for the difference between the 4sU-seq signals at each position in mock and HSV1-infected cells were calculated using Wilcoxon rank sum test and plotted as a color-coded bar below the plot. The color scheme is shown in the inset.
(B) 4sU-seq signals at genes that displayed DtoP APA shifts in HSV-1 infected cells involving intronic PAS. IPA: intronic polyadenylation site.
(C) PAS-seq and 4sU-seq tracks of TOB2. Red arrows point to the IPA region.
(D) 4sU-seq signals at genes that displayed PtoD APA shifts in HSV-1 infected cells.
(E) 4sU-seq signals at genes that displayed PtoD APA shifts in HSV-1 infected cells involving intronic PAS.
(F) PAS-seq and 4sU-seq tracks of DNAJB6.

below the plot), indicating loss of transcription activity within this region. Second, the 4sU-seq signals downstream of transcription end site (TES) in HSV-1-infected cells were higher than those in mock treated cells, consistent with DoTT (Figure 2.3A, marked by a red arrow). To better monitor potential changes in transcription activities near the regulated PAS, we focused on DtoP shifts involving proximal IPA that are used in at least 20% of the transcripts. Importantly, accumulation of 4sU-seq signals were observed at these IPA ($p < 0.05$, Wilcoxon rank sum test), followed by a decrease in HSV-1-infected cells (Figure 2.3B, a quantitative comparison for individual PAS is shown in Figure 2.4). This pattern is a hallmark of transcriptional termination (Eaton and West 2020), suggesting that the observed higher PAS-seq signals at these IPA are, at least in part, due to higher usage of these PAS during transcription. This is exemplified by the gene TOB2 (Figure 2.3C). Here, an IPA in TOB2 was activated in the HSV-1-infected cells. Concomitantly, 4sU-seq signals accumulated in this region followed by a decrease in HSV-1-infected cells, consistent with transcriptional termination. Significantly higher 4sU-seq signals were also observed downstream of the TES consistent with impaired PAS usage and read-through transcription at the canonical downstream PAS (Figure 2.3C). As a comparison, we also plotted the 4sU-seq signals in mock and HSV-1-infected cells for genes that displayed PtoD APA changes. Different from the DtoP genes (Figure 2.3A and 2.3B), the 4sU-seq signals from mock and HSV-1-infected cells were similar over the gene body for PtoD genes (Figure 2.3D), indicating that there was no early transcription termination. Additionally, the transcription termination defect was less significant (Figure 2.3D, marked by a red arrow). When examining the 4sU-seq signals at the IPA of PtoD genes, we observed the opposite pattern than the DtoP genes. A peak followed by a valley pattern was detected at these IPA in mock treated cells (Figure 2.3E, blue line), indicating polyadenylation at these sites was accompanied by

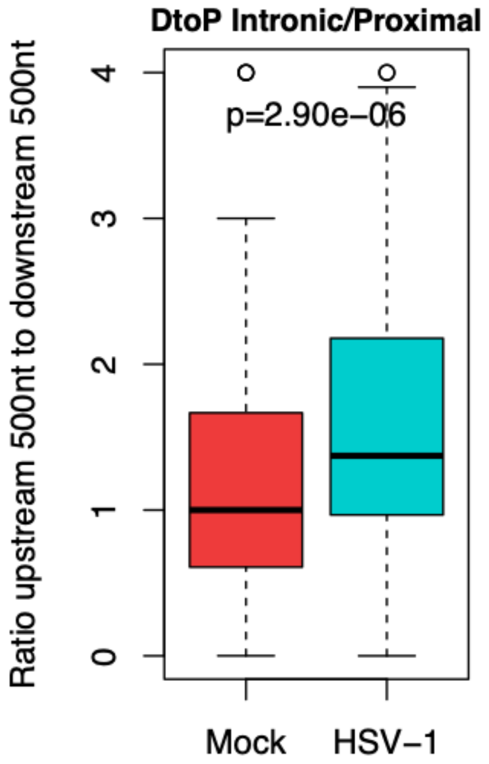

Figure 2.4 4sU-seq signal ratio between the regions 500nt upstream and 500nt downstream of the intronic PAS of DtoP genes.

transcriptional termination. By contrast, 4sU-seq signals in HSV-1-infected cells were relatively flat at these intronic PAS, indicating transcriptional read- through (Figure 2.3E, red line). This pattern is exemplified by the DNAJB6 gene (Figure 2.3F). A major decrease in 4sU-seq signals were observed at the IPA of DNAJB6 in mock-treated cells, consistent with transcription termination at this site (Figure 2.3F). In HSV-1 infected cells, transcription extended beyond the IPA, accompanied by a shift in polyadenylation to the downstream PAS (Figure 2.3F). Together, these data suggest that the HSV-1-induced changes in APA profiles are, at least in part, caused by changes in PAS usage during transcription.

ICP27-dependent and -independent APA changes during HSV-1 infection

We recently showed that the HSV-1 immediate early factor ICP27 directly interacts with the mRNA 3’ processing factor CPSF and blocks mRNA 3′ end formation (Wang et al. 2020). This suggests that ICP27 could be directly involved in HSV-1-mediated APA regulation. To test this possibility, we first compared the APA changes induced by the wild-type and the ΔICP27 HSV-1, in which the ICP27 gene was replaced by lacZ in HeLa cells (Smith et al. 1992). The majority of HSV-1-induced APA changes were abolished or diminished in ΔICP27 infected cells (Figure 2.5A, compare HSV1 and ΔICP27), strongly suggesting that ICP27 is required for HSV-1-mediated APA regulation. Interestingly, however, when comparing the APA profiles of mock and ΔICP27 infected cells, we detected 1,435 significant APA changes and the majority of these APA changes (1,109 genes or 77%) are DtoP shifts (Figure 2.5B). Therefore, the ΔICP27 virus induced an even greater number of APA changes than the wild-type virus. The APA changes induced by ΔICP27 virus seemed distinct from those by the wild-type virus. The proximal PAS

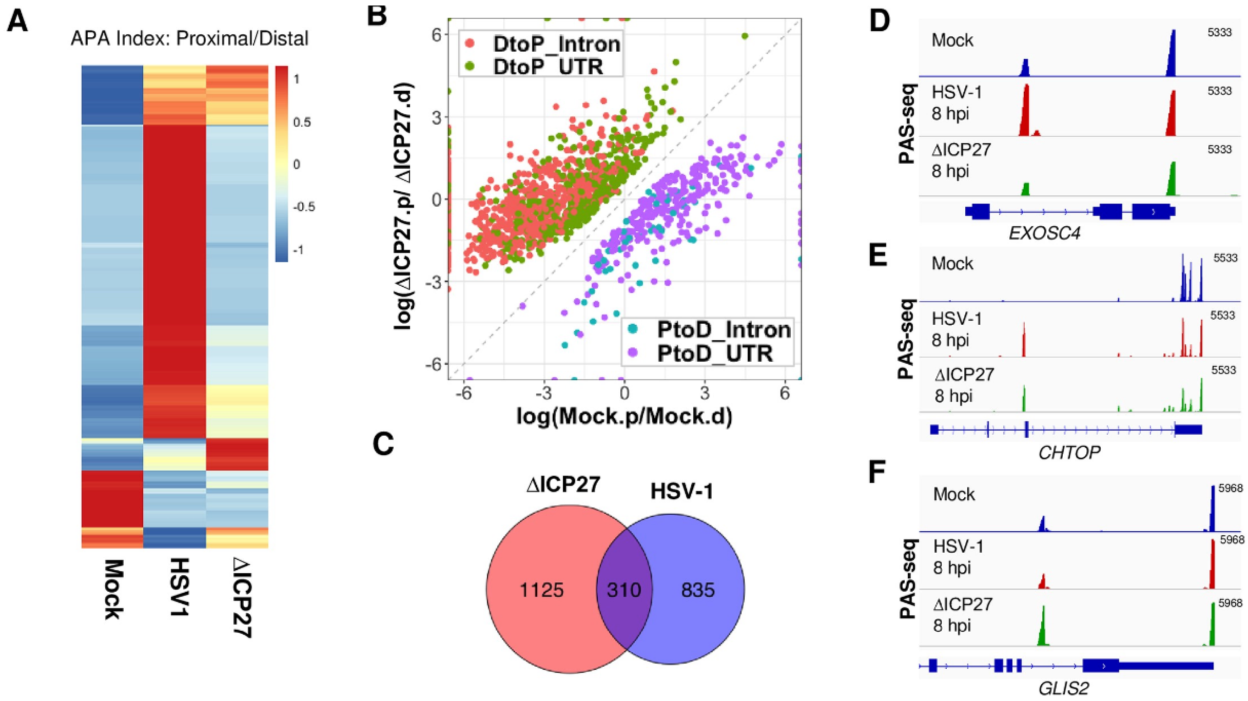

Figure 2.5 ICP27-dependent and -independent HSV-1-induced APA changes

(A) A heat map showing the APA index of all the genes that displayed significant APA changes as shown in Figure 2.1A in mock, wild-type or ΔICP27 HSV-1-infected HeLa cells (8 hpi). Data was scaled by row.
(B) A scatter plot showing significant APA differences between mock- and ΔICP27 HSV-1-infected HeLa cells (8 hpi).
(C) A Venn diagram showing the overlap between proximal PAS that displayed significant APA changes induced by the wild-type or ΔICP27 HSV-1.
(D-F) PAS-seq tracks of mock-, wild-type HSV-1, or ΔICP27 HSV-1-infected HeLa cells at 8 hpi.

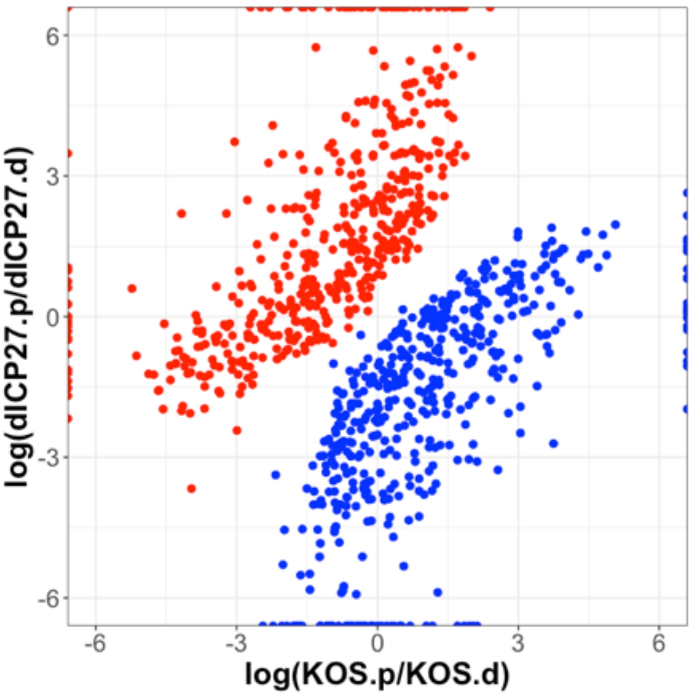

Figure 2.6 Significant APA changes between KOS and ΔICP27 HSV-1-infected cells

involved in ΔICP27- and wild-type HSV-1-induced APA changes were largely distinct with relatively small overlap (310 genes in the overlap, Figure 2.5C). Additionally, 1,000 significant APA differences were detected when we compared the wild-type and ΔICP27 HSV-1-infected cells and 501 of these genes showed more proximal PAS usage in the wild-type virus-infected cells and 409 displayed the opposite trend (Figure 2.6). These data demonstrate that HSV-1 can induce APA changes in both ICP27-dependent and -independent manners. For example, HSV-1 infection activated an intronic PAS in EXOSC4 transcripts, and a similar activation was not observed in ΔICP27-infected cells (Figure 2.5D). By contrast, an intronic PAS in CHTOP is ICP27-independent as it was similarly activated in both wild-type and ΔICP27 HSV-1-infected cells (Figure 2.5E). Finally, an intronic PAS in GLIS2 was only activated in ΔICP27- but not in wild- type HSV-1-infected cells (Figure 2.5F). Therefore, these results suggest that HSV-1 can induce APA changes through multiple mechanisms.

Mechanisms for HSV-1-mediated APA regulation

Our data suggest that ICP27 is necessary for the majority of APA changes induced by wildtype HSV-1. We thus wanted to determine if ICP27 is sufficient to regulate APA. Based on RNA-seq analyses, the Krause laboratory recently provided evidence that ectopically expressed ICP27 regulates APA (Tang et al. 2016). However, RNA-seq is not ideal for APA analysis as it lacks the sensitivity to detect APA changes of modest magnitude or those involving closely located alternative PAS (Shi 2012). To overcome these limitations, we performed PAS-seq analysis of mock-transfected or ICP27 over-expressing HEK293 cells. Overexpression of ICP27 induced significant APA changes in 169 genes, the vast majority of which (154 genes or 91%) were DtoP shifts (Figure 2.7A). Among these DtoP shifts, 111 genes or 72% shifted to a proximal intronic PAS (DtoP_Intron, Figure 2.7A). The majority of ICP27 overexpression-induced APA changes were also induced by HSV-1 infection (p < 5.1e-71, hypergeometric test; Figure 2.8A). However, the number of APA events regulated by ICP27 overexpression was significantly lower compared to that by HSV-1. Thus, although ICP27 is necessary for a majority of HSV1-induced APA changes, it is not sufficient to induce these changes.

Cleavage and polyadenylation at IPA are generally inhibited by the U1 snRNP (Kaida et al. 2010). As ICP27 overexpression primarily activates IPA, it was proposed that ICP27 may modulate APA by blocking U1 snRNP activity (Tang et al. 2016). To test this model, we transfected U1 antisense morpholino oligo (AMO) into HEK293 cells, which blocks U1 snRNA-RNA interactions and thereby inhibiting U1 activity. PAS-seq analysis showed that U1 AMO treatment resulted in significant APA changes in 1,999 genes, the majority of which (1,867 genes or 93%) were DtoP shifts (Figure 2.7B). Consistent with previous studies, the majority of these APA changes (1,646 genes or 82% of the total) involve the activation of an intronic PAS (Figure 2.7B). A comparison between U1 AMO-, HSV-1- and ICP27-induced APA changes revealed largely distinct patterns with small overlaps (Figure 2.7C). For example, among the 169 genes whose APA is regulated by ICP27, only 32 (19%) are also regulated by U1 snRNP (Figure 2.8B). Similarly, 13% of HSV1-induced APA changes were also induced by U1 AMO (Figure 2.8C). For examples, ICP27 is necessary and sufficient to activate an IPA in CCDC71 and UNC119b (Figure 2.7D). However, the intronic PAS of UNC119b, but not CCDC71, was induced by U1 AMO treatment (Figure 2.7D). These data suggest that HSV-1-mediated APA regulation does not involve a general inhibition of U1 snRNP.

To begin to understand the molecular basis for the specificity of HSV-1- and ICP27-induced activation of intronic PAS, we examined the sequences of the regulated PAS. Comparison

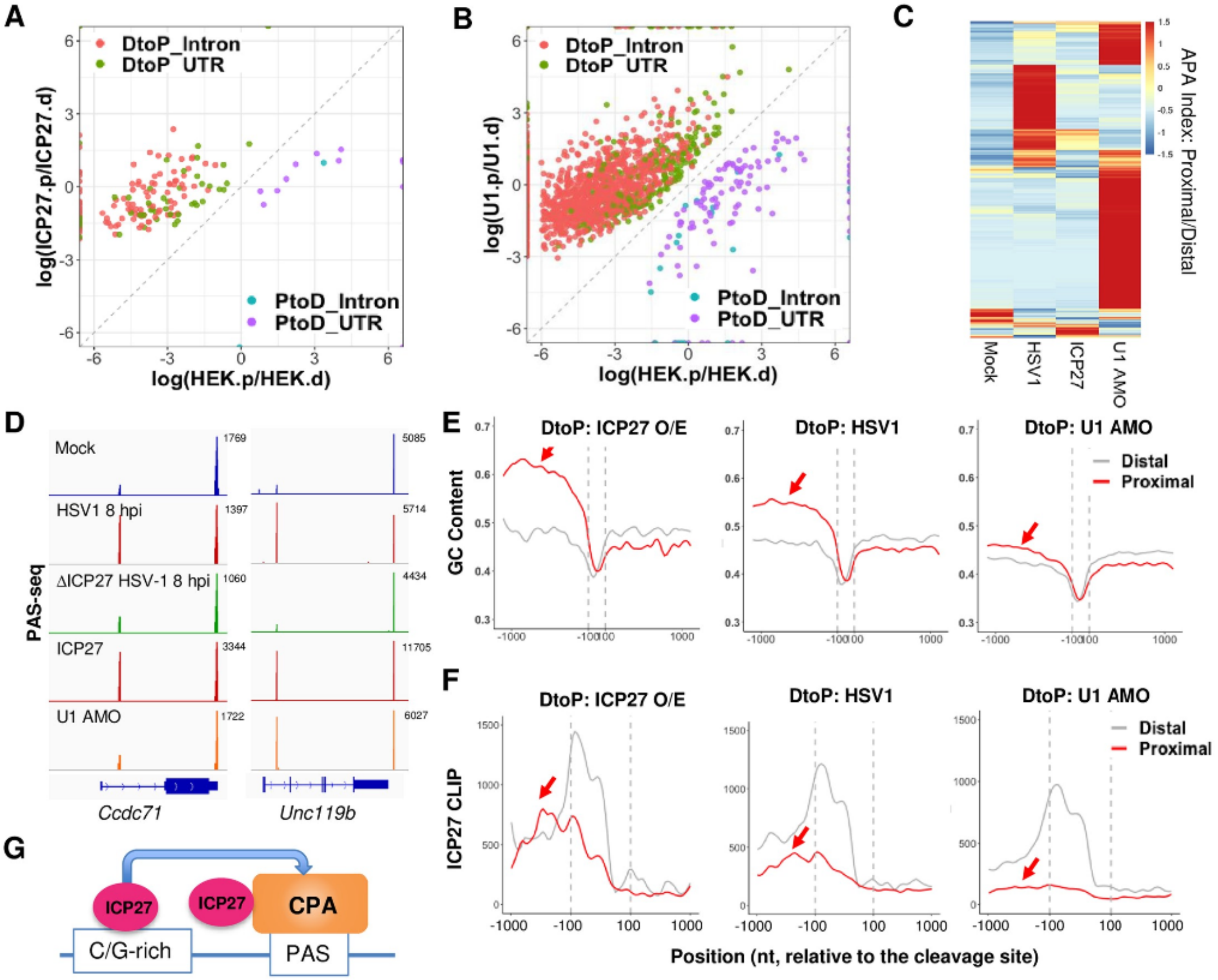

Figure 2.7. Mechanism of HSV-1-mediated APA regulation

(A) A scatter plot showing the significant APA changes induced by ICP27 over-expression in HEK293 cells. Color scheme and labeling are similar to Fig 2.1A.
(B) A scatter plot showing the significant APA changes induced by U1 antisense morpholino oligo (AMO)-treatment of HEK293 cells.
(C) A heat map of APA index of all genes that displayed significant APA changes induced by HSV-1-infection, ICP27 over-expression, or U1 AMO. Data was scaled by row.
(D) PAS-seq tracks for two example genes.
(E) GC content at the proximal and distal PAS of genes that displayed significant DtoP shifts induced by ICP27 over-expression (O/E), HSV-1 infection (HSV1), or U1 AMO treatment.
(F) ICP27 CLIP signals at the proximal and distal PAS of genes that displayed significant DtoP shifts induced by ICP27 over-expression (O/E), HSV-1 infection (HSV1), or U1 AMO treatment.

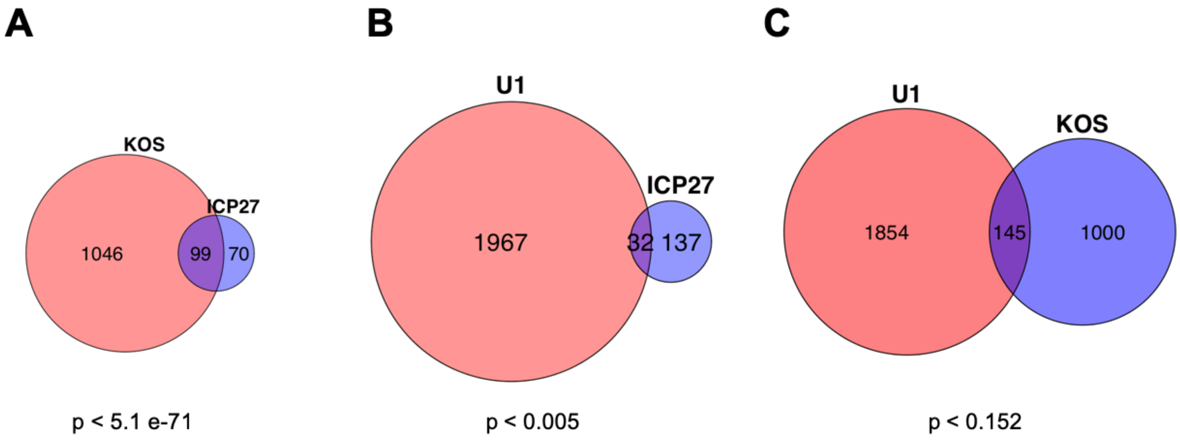

Figure 2.8 Venn diagrams showing comparison of significant APA changes in U1 AMO-treated and ICP27 over-expressing or KOS-infected cells.

P values were calculated using the hypergeometric test.

of the IPA activated by either ICP27 or HSV-1 with the corresponding distal PAS in the 3′ UTR, revealed a higher GC content at the IPA (Figure 2.7E). ICP27-activated intronic proximal PAS are highly G/C-rich in the region upstream of cleavage sites (Figure 2.7E, left panel). HSV-1-activated intronic proximal PAS show an intermediate GC content (Figure 2.7E, middle panel). By contrast, intronic PAS activated by U1 suppression had a lower GC content (less than 50%, Figure 2.7E, right panel). We have previously shown that ICP27, when bound to upstream GC-rich sequences, can activate PAS (Figure 2.7G) (Wang et al. 2020). To test if the GC contents of the different classes of IPA impact ICP27 binding, we took advantage of the ICP27 CLIP-seq dataset that we generated previously (Wang et al. 2020). Interestingly, we commonly observed high ICP27 CLIP-seq signals upstream of ICP27-activated intronic PAS and intermediate levels of ICP27 CLIP-seq signals at HSV-1-induced IPA (Figure 2.7F, left and middle panels). By contrast, very little ICP27 binding was detected upstream of U1-regulated IPA (Figure 2.7F, right panel). Thus, the ICP27 CLIP-seq signal intensities at these IPA are highly consistent with the respective GC content. These observations are consistent with the model that ICP27 activates specific PAS by binding to GC-rich upstream sequences during HSV-1 infection, and that HSV-1-mediated APA regulation does not involve a general inhibition of U1 snRNP.

Export of HSV-1-induced APA isoforms

We next wanted to determine how HSV-1-induced APA changes regulate the export of the corresponding transcripts. To address this question, we analyzed a RNA-seq dataset that we have recently generated for chromatin, nucleoplasmic, and cytoplasmic fractions of mock- or HSV-1-infected human fibroblast cells at 8 hpi (Hennig et al. 2018). Our previous study showed that known nuclear lincRNAs, including MALAT1 and NEAT1, and cytoplasmic lincRNAs, including LINC00657 and VTRNA2-1, were enriched in nuclear and cytoplasmic fractions respectively (Hennig et al. 2018). In addition, intronic reads were over-represented in chromatin-associated fraction (Hennig et al. 2018). These observations suggested that the fractionation was efficient. To measure the overall export efficiencies of HSV-1 target APA isoforms, we performed a meta-analysis of all genes that displayed significant APA changes in HSV-1-infected cells, but do not overlap with other genes within the 1 kb region downstream of TES. Compared to the RNA-seq patterns in mock treated cells (Figure 2.9A), HSV-1-infected cells displayed two major differences (Figure 2.9B). First, there was significant accumulation of RNA-seq signals downstream of TES in the chromatin and nucleoplasm fractions, consistent with HSV-1-induced DoTT. Secondly, accumulation of RNAs was observed in the nucleoplasm relative to the cytoplasm in both gene body and downstream regions (Figure 2.9B), suggesting that the transcripts that extended past the TES were released into the nucleoplasm, but not exported. The release of DoTT transcripts could be due to cleavage/polyadenylation downstream of the normal TES. Indeed, as shown in Figure 2.9C, elevated levels of PAS-seq signals were observed downstream of HNRNPA2B1 TES (PAS-seq peaks downstream of the TES are marked by red arrows). On the global level, we compared the PAS-seq reads in the 5 kilobase (kb) region downstream of the annotated TES for all PtoD genes and found that, indeed, there are significantly higher PAS-seq reads within this region in HSV-1-infected cells compared to mock treated cells (Figure 2.9D, $p = 0.002$, Wilcoxon test). The PAS-seq reads in the downstream regions contain the canonical poly(A) signal, AWTAAA hexamer, at -20 nt position (Figure 2.10A), and do not have a poly(A) run downstream of the cleavage sites (Figure 2.10B), strongly suggesting that these PAS-seq reads are due to the usage of cryptic PAS and not due to a potential technical artifact such as internal priming. This data suggests that HSV-1-induced

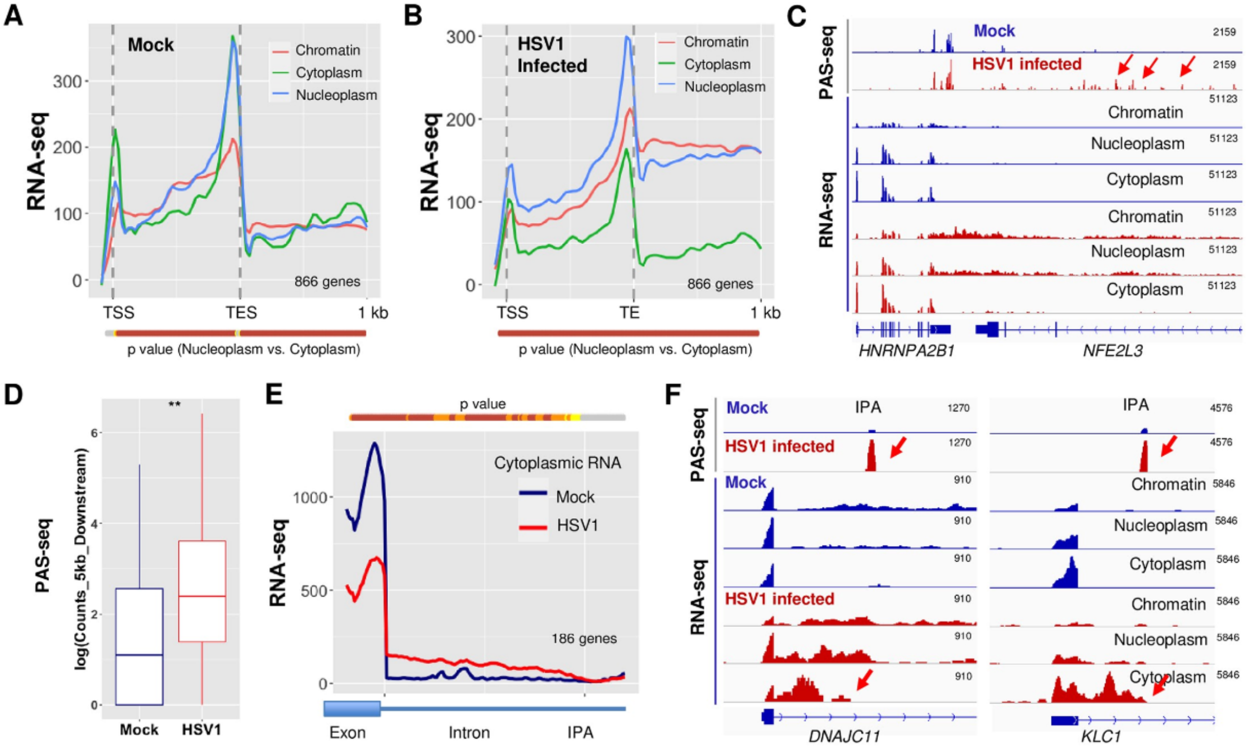

Figure 2.9. HSV-1-mediated APA regulation and mRNA export.

Average RNA-seq signals for genes that display significant HSV-1-induced APA changes in chromatin, nucleoplasm, and cytoplasm fractions in mock-infected **(A)** or HSV-1-infected human fibroblast cells (**B). ** P values for the difference in RNA-seq signals at each position between the nuclear and cytoplasmic fractions were calculated using Wilcoxon rank sum test and plotted as a color-coded bar below the plot. The color scheme is the same as Figure 2.3A.
(C) PAS-seq and RNA-seq tracks for HNRNPA2B1 gene. Red arrow points to the PAS-seq signals downstream of the normal TES.
(D) PAS-seq signals in the 5kb region downstream of the normal TES in mock- and HSV-1-infected cells. ** :P value < 0.05, Wilcox test.
(E) Cytoplasmic RNA-seq signals for HSV-1-induced DtoP_intron APA changes. IPA: intronic poly(A) site.
(F) PAS-seq and RNA-seq tracks for two example genes. Red arrows point to the IPA.

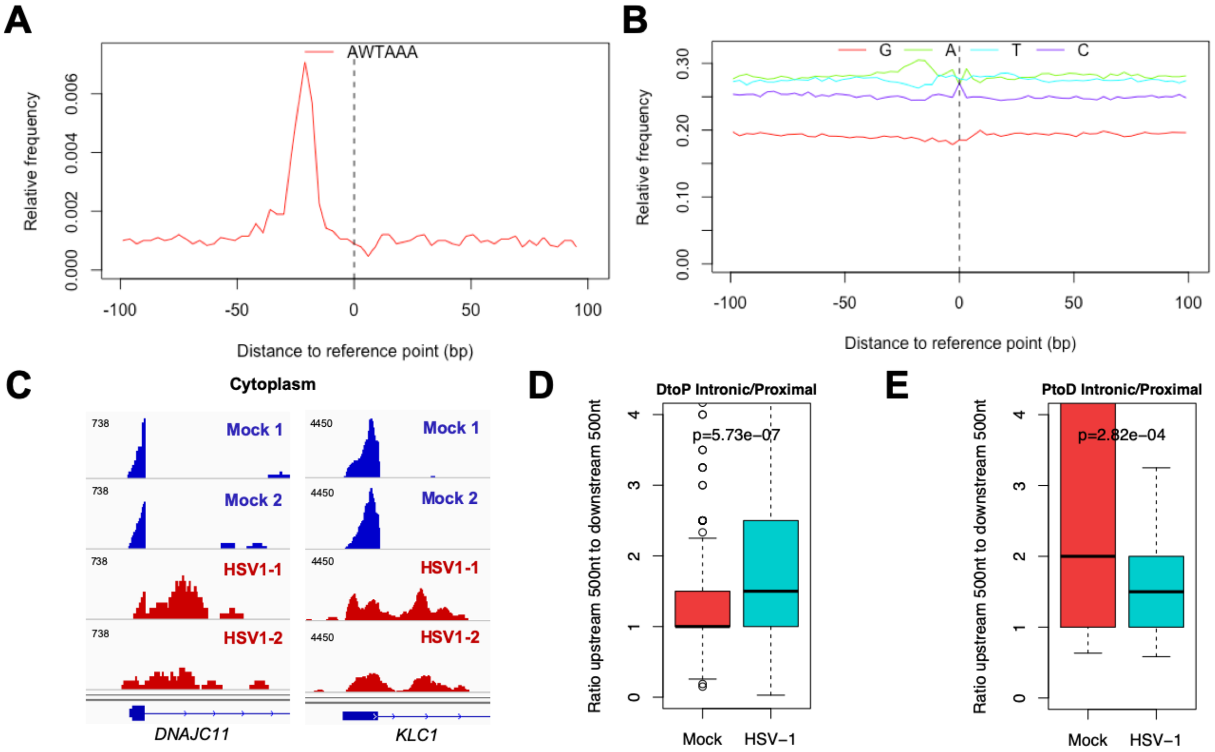

Figure 2.10 DoTT transcripts are processed at cryptic PAS

(A) Distribution of AWTAAA motif at the PAS-seq peaks downstream of normal TES.

(B) Nucleotide composition of cryptic PAS downstream of normal PAS.

(C) RNA-seq tracks of cytoplasmic RNAs in mock- and HSV-1-infected cells (8 hpi).

(D)-(E) The ratio of RNA-seq read counts in the 500 nt upstream and 500 nt downstream of the intronic PAS for DtoP **(D)** and PtoD **(E)** genes for the cytoplasmic fractions of mock- and HSV-1-infected cells.

extended transcripts as a result of DoTT are released into the nucleoplasm by cleavage/polyadenylation at cryptic PAS in the downstream region. However, these transcripts are not efficiently exported into the cytoplasm (Figure 2.9B and 2.9C).

HSV-1 infection activates IPA in a large number of genes (Figure 2.1A). The resultant transcripts are predicted to encode truncated proteins. To monitor the fate of these RNAs, we performed a meta-analysis of the region from the upstream exon to the intronic PAS for DtoP_intronic genes, which distinguishes the spliced and polyadenylated APA isoforms (Figure 2.9E). Signals from the upstream exon reflect both spliced and polyadenylated transcripts whereas the signals in the intronic region are only derived from the unspliced polyadenylated isoform. In the cytoplasm of mock infected cells, high RNA-seq signals were observed for the upstream exon while almost no signals were detected in the intronic regions (Figure 2.9E, blue line), suggesting that only fully spliced transcripts are exported. However, RNA-seq signals decreased in the upstream exon region, but accumulated between the upstream exon and the IPA in the cytoplasm of HSV-1-infected cells (Figure 2.9E, red line). This suggests that the transcripts polyadenylated at IPA are exported into the cytoplasm. Two examples were provided in Figure 2.9F. For both DNAJC11 and KLC1, their transcripts are efficiently spliced in mock treated cells, but HSV-1 infection activates a PAS within the first intron, as shown by the PAS-seq data (Figure 2.9F, PAS-seq tracks, activated IPA are marked by red arrows). Our fractionation RNA-seq data showed that these truncated RNA isoforms are exported into the cytoplasm (Figure 2.9F, RNA-seq tracks, cytoplasmic tracks are marked by red arrows). These results are highly consistent between the two biological replicates of sub-cellular fractionation in our dataset (Figure 2.10C). Based on these observations, we conclude that the transcripts of the APA target genes are exported less efficiently and that the truncated transcripts polyadenylated at intronic PAS are exported. We further estimated the IPA isoform export efficiency by calculating the ratio of RNA-seq signals within the 500 nt region upstream of the intronic PAS to the 500 nt region downstream in all APA changes involving proximal IPA (Figure 2.10D and 2.10E). The results suggest that the intronic reads upstream of the IPA increased for DtoP genes (Figure 2.10C), but decreased in PtoD genes (Figure 2.10D), further suggesting that the IPA isoforms are exported.

Translation of HSV-1-induced APA isoforms

Our data suggests that at least a subset of the HSV-1-induced APA isoforms polyadenylated at IPA are exported into the cytoplasm (Figure 2.9E and 2.9F and 2.10C and 2.10D), raising the question whether they are translated. To test this, we examined our Ribo-seq dataset for HSV-1-infected human fibroblast cells at different time points post-infection. In un-infected cells, the Ribo-seq signals were limited to exonic regions as expected (Figure 2.11A). Interestingly, however, Ribo-seq signals extended into intronic regions within many of the HSV-1-induced APA isoforms (Figure 2.11A). These results not only further support our conclusion that intronically polyadenylated APA isoforms can be exported into the cytoplasm, but also indicate that they are engaged with ribosomes. For 54 out of 132 intronically polyadenylated APA isoforms that had least 5 intronic codons upstream of the first intronic stop codon and where the upstream exon was translated, the intronic read density exceeded 5% of the upstream exon at 4–8 hpi. Several lines of evidence indicate that these reads correspond to ribosomes that continue translation elongation from the upstream exon into the intron: 1) Intronic translation was much weaker in un- infected samples and during early infection (before 4 hpi) (Figure 2.11B; $p = 3.1x10^{-4}$, Kolmogorov-Smirnov test); 2) Mapping Ribo-seq reads to nucleotide resolved ribosome positions (Erhard et al. 2018) revealed a strong enrichment of in-frame codons, providing strong evidence for actively translating ribosomes (Figure 2.11C); 3) Virtually no reads were observed downstream of the first intronic in-frame stop codon (Figure 2.11C); 4) Lactimidomycin (ltm) or harringtonine (harr) treated samples, in which translation is stalled at the initiation stage (Erhard et al. 2018), exhibited lower intronic read densities, suggesting that the Ribo-seq signal is not due to spurious intronic translation initiation (Figure 2.11C and Figure 2.12). Taken together, these results provide strong evidence that at least some of the HSV-1-induced IPA isoforms are translated at considerable levels.

2.5 Discussion

mRNA 3′ end processing and transcription termination are tightly coupled processes. Viral infections (HSV-1 and IAV) and cellular stresses (salt/osmotic stress and heat shock) induce DoTT/DoG (Rutkowski et al. 2015; Zhao et al. 2018; Vilborg et al. 2015). Meanwhile, several pathogens, including viruses (HCMV and VSV) and bacteria (listeria and salmonella), as well as arsenic stress causes widespread APA changes (Batra et al. 2016; Jia et al. 2017; Zheng et al. 2018; Pai et al. 2016). However, no study has characterized these two processes in response to the same pathogen or stress. In this report, we performed extensive transcriptomic analyses of wild-type and mutant HSV-1-infected cells and found that lytic HSV-1 infection induced widespread APA changes in host transcripts, the majority of which shifted to upstream PAS. HSV-1-mediated APA regulation requires the viral immediate early factor ICP27 as well as other viral factors, but does not involve a general inhibition of U1 snRNP. Interestingly, HSV-1 induces both activation of upstream PAS with pre-mature transcription termination and a

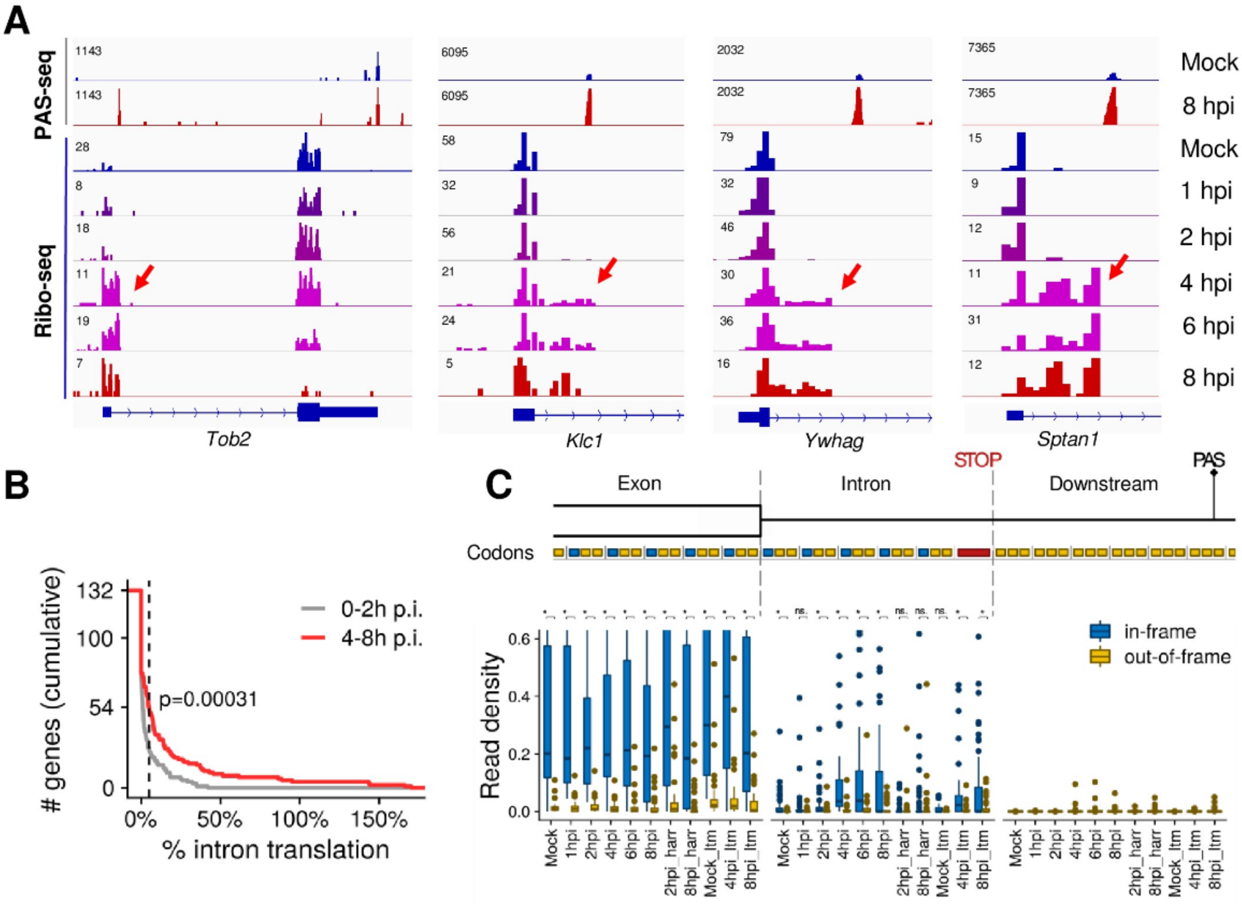

Figure 2.11 HSV-1-induced intronically polyadenylated APA isoforms are translated.

(A) PAS-seq and Ribo-seq tracks for four example genes. Ribo-seq signals within intronic regions are marked by red arrows.
(B) Cumulative distribution of the percentage of intronic Ribo-seq read densities (normalized reads per codon) compared to the level in the upstream exon. The dashed line indicates genes exceeding the 5% threshold mentioned in the text. The P value for comparing the pooled read densities in un-infected through 2 hpi vs. 4–8 hpi is shown (two-sided Kolmogorov-Smirnov test)
(C) Boxplots showing the distributions of read densities for the 54 genes exceeding the 5% threshold stratified by time point after infection, location with respect to exon-intron boundary and first intronic stop codon, and reading frame of translation. The hinges and whiskers correspond to quartiles and to the most extreme values outside of 1.5 times the inter-quartile range, respectively. The median and outliers are indicated. The y axis is arbitrarily cut at 0.6. P values for comparisons of in-frame and out-of-frame codons are indicated (***, p<0.001; **, p<0.01; *, p<0.05; n.s., not significant at 5% level).

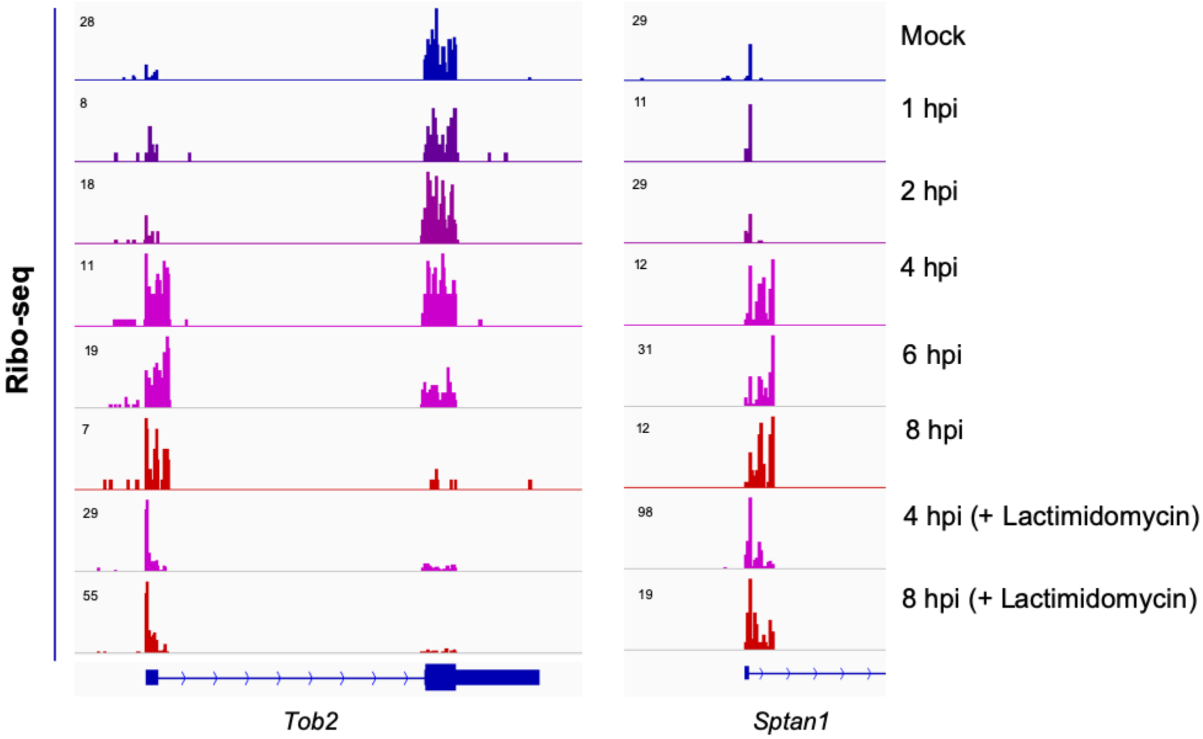

Figure 2.12 Ribo-seq example genes

Lactimidomycin is a translation inhibitor.

termination defect. Activation of upstream intronic PAS produces truncated transcripts that are exported into the cytoplasm and translated. By contrast, although extended transcripts due to DoTT can be cleaved and polyadenylated at downstream cryptic PAS, these transcripts are sequestered in the nucleoplasm. Together, these results demonstrate that HSV-1-mediated regulation of APA and transcription termination profoundly reprograms host transcriptomes (Figure 2.13).

Although widespread APA changes have been described for a number of pathogen-infected cells and for cells exposed to arsenic stress (Batra et al. 2016; Jia et al. 2017; Zheng et al. 2018; Pai et al. 2016), the underlying mechanism remains poorly understood. Our data suggest that HSV-1 induces APA changes of host mRNAs through multiple mechanisms. First, the viral immediate early factor ICP27 contributes to HSV-1-induced APA changes. We have previously shown that ICP27 has bimodal activities: it broadly inhibits mRNA 3′ processing through direct interactions with the 3′ processing factor CPSF, but can activate PAS that contain GC-rich upstream sequences (Wang et al. 2020). Indeed, both HSV- 1 infection and over-expression of ICP27 can activate upstream intronic PAS and these PAS contains GC-rich upstream sequences (Figure 2.7E and 2.7F). 3′ processing at these intronic PAS induces early termination (Figure 2.3A–2.3C). On the other hand, the corresponding downstream PAS in these genes lack GC-rich upstream sequences and are thus inhibited, leading to DoTT at these sites. Thus, the bimodal activities of ICP27 provide an explanation for the paradoxical observation of early termination and DoTT in these genes. Secondly, our observation that the ΔICP27 virus still induce a large number of APA. Similarly, we have previously shown that ΔICP27 HSV-1 also induced DoTT, albeit at lower levels compared to that by the wild-type virus (Wang et al. 2020). These results suggest that other mechanisms are also involved. Although ICP27 is required for

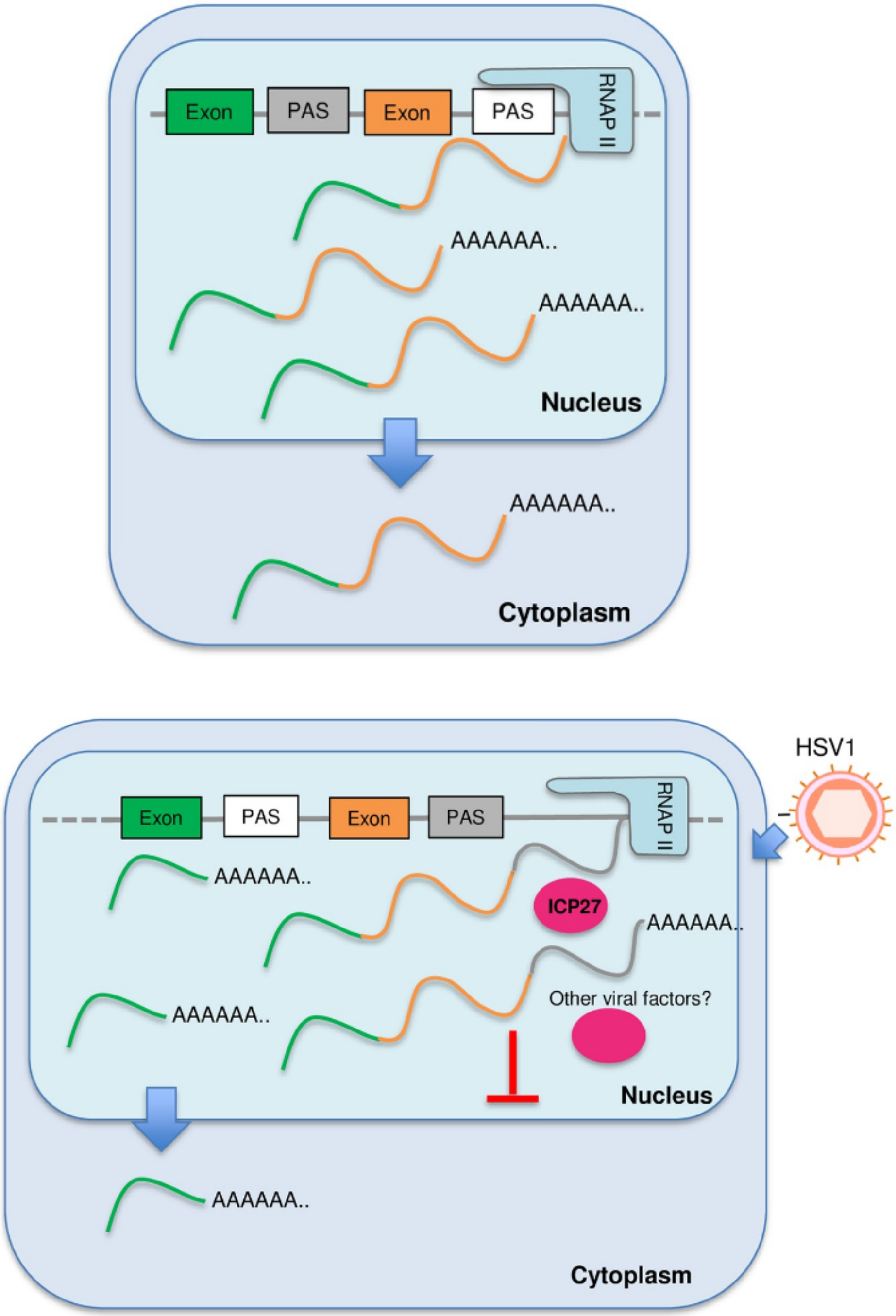

Figure 2.13. A model for HSV-1-mediated APA regulation.

In HSV-1-infected cells, ICP27 and other viral factors induce many APA changes and transcription termination defects. Transcripts that extend beyond the normal TES are cleaved and polyadenylated, but are not exported. Truncated transcripts that are polyadenylated at IPA can be exported into the cytoplasm and translated. Please see text for more details.

viral replication and for the expression of early and late genes, the ΔICP27 virus still contains the tegument proteins VP16 and vhs, and other immediate early proteins, such as ICP4, ICP0 and ICP22 are also expressed (Sandri-Goldin 2011). Thus the viral DNA and other viral proteins may induce APA changes and DoTT either directly through interactions with host mRNA 3' processing factors or indirectly. Since multiple pathogens and stress induce similar changes in both APA and DoTT, it is likely that a common mechanism underlies these phenomena. One possibility is that viral infections and cellular stress may alter the activity of RNAPII. In addition to its role in transcribing genes, RNAPII also plays an essential role in coordinating transcription and RNA processing primarily through its C-terminal domain. Both phosphorylation and dephosphorylation of RNAPII CTD have been shown to influence termination (Richard and Manley 2009; Eaton and West 2020; Proudfoot 2016; Johnson et al. 2009). For example, pharmacological or genetic inhibition of Cdk12, which phosphorylates RNAPII CTD at serine 2, leads to activation of intronic PAS and premature termination (Dubbury et al. 2018; Krajewska et al. 2019). On the other hand, PP1 or PP2A, phosphatases that dephosphorylate RNAPII CTD, play essential roles in regulating transcription pausing and termination (Eaton and West 2020; Cortazar et al. 2019; Kecman et al. 2018; Huang et al. 2020). Previous studies provided evidence that HSV-1 infection induces aberrant CTD phosphorylation and partial degradation of RNAPII (Dai-Ju et al. 2006; Fraser and Rice 2007). It will be important to characterize RNAPII post-translational modifications and interactomes in pathogen-infected and in stressed cells and determine if/how such changes contribute to the virus-induced APA changes and DoTT.

The functional consequence of pathogen/stress-induced DoTT and APA changes remains unclear. The most important functions of stress responses are to: 1) shut down the expression of most genes to avoid accumulation of aberrant proteins; 2) activate stress response genes to stabilize and repair biomolecules (Galluzzi et al. 2018). Similarly, when a pathogen infects a host cell, it shuts down host gene expression and hijacks the host machinery to express genes of the pathogen. Both DoTT and APA changes could contribute to the repression of cellular genes. DoTT inter- feres with the transcription cycle and prevents mRNA biogenesis. Consistent with previous reports (Hennig et al. 2018), our results showed that at least some of the read-through transcripts as a result of DoTT are in fact cleaved and polyadenylated, and released into the nucleoplasm (Figure 2.9). However, they are not efficiently exported. On the other hand, HSV-1-induced activation of upstream intronic PAS leads to the production of truncated transcripts that can be exported (Figure 2.9E). We provide evidence that at least some these truncated mRNAs are translated (Figure 2.11). Therefore, both DoTT and APA may function in host shutoff (for pathogens) or repressing bulk gene expression (for stresses). Alternatively, the DoTT and APA changes observed in pathogen-infected or stressed cells could represent a host defense mechanism. Previous studies provided evidence that arsenic stress-induced APA isoforms with shorter 3' UTRs, which can evade RNA degradation, are thus better preserved (Zheng et al. 2018). This may facilitate better recovery from stress. VSV-induced APA changes have been shown to modulate the innate immunity response (Jia et al. 2017). In summary, pathogen- and stress-induced APA changes may function in host shut-off or in host defense, and these two mechanisms are not mutually exclusive.

2.6 Methods:

Cell culture, viruses and infection

HEK293 and HeLa cell lines were cultured in Dulbecco's modified Eagle medium (DMEM) with 10% fetal bovine serum (FBS). All cells were incubated at 37℃ in a 5% (v/v) CO2- enriched incubator. Virus stocks for wild-type HSV-1 strain KOS as well as the ICP27 null mutant (strain KOS) (Smith et al. 1992) were produced on complementing Vero 2–2 cells (Sekulovich et al. 1988). HeLa cells were infected with an MOI of 10 unless otherwise specified and incubated at 37℃ until cells were harvested at the specified time points. For anti-sense morpholino oligo treatment, HEK293 cells were treated with 50 μM U1 antisense morpholino oligo (AMO) (Gene tools) and 10 μM Endo-Porter (Gene tools). After 48 hours, RNA was extracted by using Trizol (Ambion). ICP27 over-expression was also performed in HEK293 cells.

PAS-seq

Total RNA was extracted with Trizol as per manual (Life technologies), 10 μg total RNA was fragmented with fragmentation reagent (Ambion) at 70℃ for 10 minutes followed by precipi- tation with ethanol. After centrifugation, RNA was dissolved and Reverse transcription was performed with PASSEQ7-2 RT oligo: [phos]NNNNAGATCGGAAGAGCGTCGTGT TCGGATCCATTAGGATCCGAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTT TTT[V-Q] and Superscript III. cDNA was recovered by ethanol precipitation and centrifugation. 120–200 nucleotides of cDNA was gel-purified and eluted from 8% Urea-PAGE. Recovered cDNA was circularized with CircligaseTM II (Epicentre) at 60℃ overnight. Buffer E

(Promega) was added in cDNA and heated at 95℃ for 2 minutes, and then cool to 37℃ slowly. Circularized cDNA was linearized by adding BamH I (Promega). cDNA was centrifugated after ethanol precipitation. PCR was carried out with primers PE1.0 and PE2.0 contain- ing index. Around 200 base pairs of PCR products was gel-purified and submitted for sequencing (single read 100 nucleotides). PAS-seq samples include: HSV-1-infected HeLa cells at 0, 2, 6, and 12 hpi (one for each time point); mock-, wild-type HSV-1, or ΔICP27 HSV- 1-infected HeLa cells at 8 hpi (one for each); mock-, ICP27 over-expressing, and U1 AMO- treated HEK293 cells (one for each). For APA analysis, mock-treated and HSV-1-infected HeLa cells at 0 hpi were considered as biological replicates and HSV-1-infected HeLa cells at 6 and 8 hpi as replicates.

PAS-Seq data analysis

From the raw PAS-seq reads, first those with no poly(A) tail (less than 15 consecutive "A"s) were filtered out. The rest were trimmed and mapped to hg19 genome using STAR. If 6 conse- cutive "A"s or more than 7 "A"s were observed in the 10 nucleotide downstream of PAS for a reported alignment, it was marked as a possible internal priming event and removed. The big- wig files were then generated for the remaining reads using deepTools (v2.4) with "normalizeUsingRPKM" and "ignoreDuplicates" parameters (Ramírez et al. 2014).

Next, the locations of 3' ends of the aligned reads were extracted and those in 40nt of each other were merged into one to provide a list of potential PAS for human. This list was then annotated based on the canonical transcripts for known genes. The final read count table was created using the reads with their 3′ ends in -40nt to 40nt of these potential PAS.

Alternatively polyadenylated PAS in different experimental conditions were identified using diffSpliceDGE and topSpliceDGE from edgeR package(v3.8.5) (Robinson et al. 2010).

This pipeline first models the PAS read counts for all PAS, then compares the log fold change of each PAS to the log fold change of the entire gene. This way, these functions, primarily used to find differential exon usage, generate a list of sites with significant difference between our PAS-seq samples. From this list, those with a FDR value less than 0.05 and more than 15% difference in the ratio of PAS read counts to gene read counts (normalized by sequencing depth) between samples were kept, and finally for each gene the top two were chosen based on P-value and marked dis- tal or proximal based on their relative location on the gene. For PAS-seq comparisons without replicates, Fisher's exact test was used to compare read counts at a PAS and the total read counts from the same gene. The P values were adjusted by the Benjamini–Hochberg method for calculating the FDR.

For the genes with alternatively polyadenylated sites (target genes), the log2 of ratio of read counts in the distal site to the read counts in the proximal site was calculated and illustrated as a heatmap in Figs 1C, 2A and 3C with pheatmap in R. The heatmap is hierarchically clustered using Pearson correlation of the gene profiles in different experiments.

Ribo-seq analysis

We applied Bowtie 1.0 (REF) to map reads to rRNA, genomic and transcriptomic sequences from the Ensembl database (version 75). rRNA reads and reads mapping to the mitochondrial genome were discarded. All alignments were mapped to genomic coordinates. Fractional counts were used for ambiguous alignments (with regard to genomic coordinates). We then used the probabilistic model implemented in Price (version 1.0.3b) (Erhard et al. 2018) to map reads to their P site codons using default parameters. All read counts corresponding to translation start site profiling (lactimidomycin-treated samples) were discarded. Next, we removed (i) PAS that were located inside of Ensembl 75 exons (n = 110), (ii) PAS without an annotated or Price- identified open reading frame (ORF) in the upstream exon (n = 64), (iii) PAS without an in- frame stop codon in the intron in between the exon boundary and the PAS (n = 9), (iv) PAS where the first in-frame stop codon was in the first five intronic codon triplets (n = 100), and (v) PAS downstream of very weakly translated ORFs (<0.5 reads per exonic codon in all Ribo- seq samples pooled; n = 104). For the remaining n = 132 PAS, we counted codon mapped reads for the partial open reading frame in the upstream exon (reads mapped to codons in the same frame as the ORF, and in the other two frames), for its extension into the intron up to the first in-frame stop codon (reads mapped to codons in the same frame as the ORF, and in the other two frames), and reads mapped in between the stop codon and the PAS (Figure 2.11C). Read counts were normalized to the total number of Ribo-seq reads mapped to the human genome.

Meta-analysis

Meta-analyses of read distribution were performed using deeptools [39]. 4SU-seq, iCLIP-seq, or RNA-seq reads were first mapped to the human genome (hg19), and then normalized by library size to produce bigwig files using the bamCoverage tool in deepTools. Variable sized regions (gene body or the region between TSS and IPA) were divided into 100 bins. Fixed sized regions were divided into 10 nt bins. Sequencing signal scores for each bin were calculated using deepTools. For meta-analyses in Figure 2.3, signal scores for each gene were further normalized by their sum before calculating the average scores. To evaluate the statistical significance of the meta-analysis results in Figures 2.3 and 2.9, p values were calculated for the sequencing signal scores of all the genes in mock and HSV-1 samples at each nucleotide position using the

Wilcoxon rank sum test, and the results are showed as color-coded bars under each plot. For the analyses shown in Figures 2.3 and 2.9A and 2.9B, additional filtering was performed to remove the genes that overlap with other genes within the 1kb downstream region. For the analysis in Figure 2.9E, DtoP genes were filtered to keep only those whose first intronic PAS was activated by HSV-1 infection to minimize the influence of different annotations of upstream exons.

Data and software availability

RNA-seq data on the subcellular RNA fractions, 4sU-seq, and Ribo-seq data were previously published (Wang et al. 2020; Hennig et al. 2018). PAS-seq data have been deposited to the GEO database (GSE151104).

CHAPTER 3

The anti-cancer compound JTE-607 reveals hidden sequence specificity of the mRNA 3′ processing machinery

3.1 Summary

JTE-607 is a small molecule compound with anti-inflammation and anti-cancer activities. Upon entering the cell, it is hydrolyzed to Compound 2, which directly binds to and inhibits CPSF73, the endonuclease for the cleavage step in pre-mRNA 3′ processing. Although CPSF73 is universally required for mRNA 3′ end formation, we have unexpectedly found that Compound 2-mediated inhibition of pre-mRNA 3′ processing is sequence-specific and that the sequences flanking the cleavage site (CS) are a major determinant for drug sensitivity. By using massively parallel in vitro assays, we have measured the Compound 2 sensitivities of over 260,000 sequence variants and identified key sequence features that determine drug sensitivity. A machine learning model trained on these data can predict the impact of JTE-607 on poly(A) site (PAS) selection and transcription termination genome-wide. We propose a biochemical model in which CPSF73 and other mRNA 3′ processing factors bind to RNA of the CS region in a sequence-specific manner and the affinity of such interaction determines the Compound 2 sensitivity of a PAS. Together, our study not only characterized the mechanism of action of a compound with clinical implications, but also revealed a previously unknown sequence-specificity of the mRNA 3′ processing machinery.

3.2 Introduction

Almost all eukaryotic mRNA 3′ end are formed through two chemical reactions, an endonucleolytic cleavage followed by polyadenylation (Colgan and Manley 1997; Chan et al. 2011). Pre-mRNA 3′ processing is not only an essential step in gene expression, but also an important mechanism for gene regulation. ~70% of human genes produce multiple mRNA isoforms by selecting different poly(A) sites (PASs), a phenomenon called alternative polyadenylation (APA) (Shi 2012; Mitschka and Mayr 2022; Tian and Manley 2016). Distinct APA isoforms from the same gene can produce functionally different proteins and/or they can be regulated differently. APA is regulated in a developmental stage- and tissue-specific manner and mis-regulation of APA contributes to many human diseases. It remains poorly understood how APA is regulated in physiological or pathological contexts and pharmacological tools are needed for manipulating APA for research and therapeutic purposes.

The sites for canonical mRNA 3′ processing, or PASs, are defined by several cis-elements, including the AAUAAA hexamer, the U/GU-rich downstream elements, and other auxiliary sequences (Colgan and Manley 1997; Chan et al. 2011). These cis-elements are recognized by multiple trans acting factors, including cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF), which in turn recruit other mRNA 3′ processing factors to assemble the pre-mRNA 3′ processing complex. Pre-mRNA cleavage is carried out by the endonuclease CPSF73(Mandel et al. 2006), which, together with CPSF100 and symplekin, forms the nuclease module of the CPSF complex mCF (Shi and Manley 2015). CPSF73 preferentially cleaves after CA or UA sequences (Sheets et al. 1990). Although the sequences flanking the CS display distinct and well-conserved nucleotide composition patterns (Ozsolak et al. 2010; Liu et al. 2017; Derti et al. 2012). it remains unknown what role, if any, these sequences play in pre-mRNA 3′ processing.

In recent years, CPSF73 has emerged as a drug target for treating a variety of diseases. For example, a number of small molecule drugs for treating toxoplasma gondii (causes toxoplasmosis) (Palencia et al. 2017), African trypanosomes (causes sleeping sickness) (Begolo et al. 2018), and Plasmodium (causes malaria) (Sonoiki et al. 2017), target the CPSF73 homologues in these pathogens. JTE-607 is a small molecule that inhibits the production of multiple cytokines by mammalian cells (Sasaki et al. 2003; Uesato et al. 2006; Jian et al. 2004). Animal studies demonstrated that administration of JTE-607 results in improvements in several inflammation diseases, including septic shock, acute injury, and endotoxemia (Sasaki et al. 2003; Uesato et al. 2006; Jian et al. 2004). Furthermore, JTE-607 was recently shown to have anti-cancer activities and specifically kill myeloid leukemia and Ewing’s sarcoma cells (Kakegawa et al. 2019; Ross et al. 2020). JTE-607 is a prodrug and is hydrolyzed to Compound 2 upon entering the cells by the cellular enzyme CES1 (Kakegawa et al. 2019; Ross et al. 2020). Compound 2 specifically binds to CPSF73 near its active site to inhibit its activity. In addition to its potential clinical application, JTE-607 has quickly become an important tool for research (Gutierrez et al. 2021; Boreikaite et al. 2022). However, it is unclear if all pre-mRNA 3′ processing events in the human transcriptome are equally affected by JTE-607 and it is unclear why this compound is only active against specific cancer types.

Although the JTE-607 target, CPSF73, is universally required for pre-mRNA 3′ processing, we have found, surprisingly, that JTE-607-mediated inhibition of pre-mRNA 3′ processing is sequence-specific both in vitro and in cells. We have identified the CS region as a major determinant of drug sensitivity. Using massively parallel in vitro assay (MPIVA) coupled with machine learning, we have comprehensively characterized the relationship between the CS sequence and JTE-607 sensitivity and identified key sequence features that determine drug sensitivity. Using the MPIVA data, we trained a machine learning model, C3PO, that can accurately predict JTE-607 sensitivity of a PAS based on its CS region sequence. We demonstrated that C3PO can predict the effect of JTE-607 on PAS selection and transcription termination genome-wide. Together, our study not only better characterized the properties of an anti-cancer and anti-inflammation compound, but also revealed a previously unknown sequence-specificity of the mRNA 3′ processing machinery.

3.3 Results

Compound 2-mediated inhibition of mRNA 3′ processing in vitro is sequence-dependent

To better understand the mechanism of action for Compound 2, the active form of JTE-607 (Ross et al. 2020). we characterized its effect on pre-mRNA processing in an in vitro cleavage assay using HeLa cell nuclear extract (NE). We first performed in vitro cleavage assays with L3, the PAS of the adenovirus major late transcript, in the presence of DMSO or increasing concentration of Compound 2 (0.1, 0.5, 2.5, 12.5, 62.5, and 100 μM). Our results showed that the cleavage of L3 PAS was strongly inhibited by Compound 2 with an IC_{50} (concentration needed to achieve 50% of maximal inhibition) of 0.8 μM (Figure 3.1A). Compound 2-mediated inhibition of pre-mRNA cleavage could occur at the cleavage step and/or the earlier pre-mRNA 3′ processing complex assembly step. To distinguish between these possibilities, we monitored pre-mRNA 3′ processing complex assembly on L3 PAS in the presence of DMSO or increasing concentrations of Compound 2 using an electrophoretic mobility shift assay. The pre-mRNA 3′ processing complex assembled indistinguishably under all conditions tested (Figure 3.1B). These results suggest that Compound 2 does not interfere with pre-mRNA 3′ processing complex assembly, but blocks cleavage of the L3 PAS.

We next performed similar in vitro cleavage assays on other PASs. Surprisingly, we found that different PASs displayed different sensitivities to Compound 2-mediated inhibition of pre-mRNA cleavage. For example, significant cleavage was observed for SVL, the PAS from SV40 late transcript, even at the highest concentration tested of Compound 2 (Figure 3.1C). The estimated IC_{50} for SVL PAS was greater than 100.2 μM (Figure 3.1C). Therefore, the IC_{50} of L3 and SVL PASs differ by over 100-fold. Similar to L3, mRNA 3′ processing complex assembly on SVL PAS was not affected by Compound 2 (Figure 3.2). In total we performed the same in vitro cleavage assays with 38 different PASs and found that their IC_{50} values varied widely (Figure 3.1D). To begin to understand the molecular basis for such variations, we first asked if the Compound 2 sensitivity of a PAS is determined by its strength, i.e. the efficiency by which it is processed by the pre-mRNA 3′ processing machinery. We measured the percentage of pre-mRNA cleaved in vitro in the absence of the drug and compared this value with their IC_{50}. Our results detected poor correlation between the two measurements (r=0.38) (Figure 3.1D, Table 3.1). We conclude that the cleavage of different PASs display differential sensitivity to Compound 2 in vitro and that the sensitivity of a PAS is not determined by its strength.

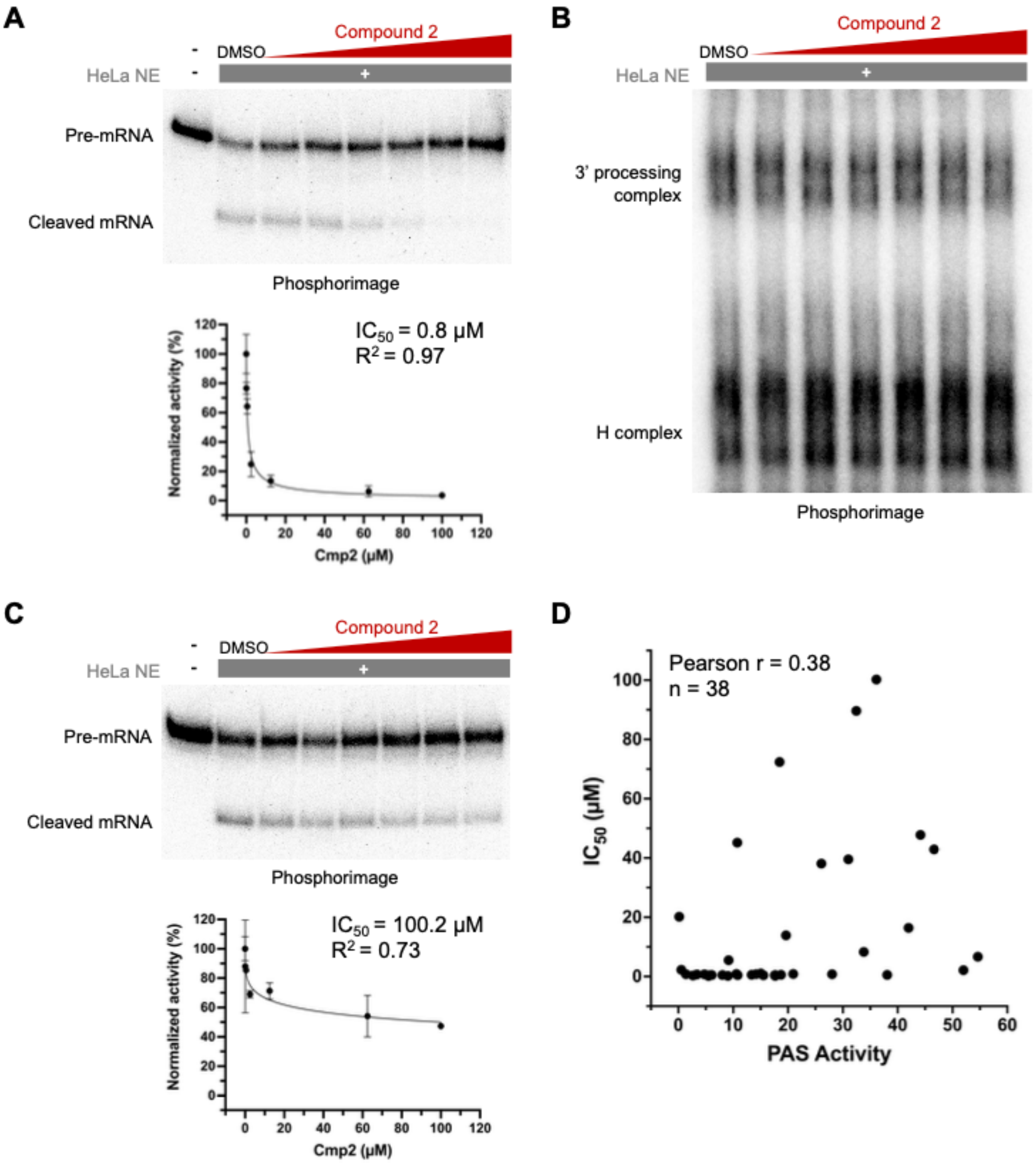
A
Compound 2
DMSO
HeLa NE
-
-
+
Pre-mRNA
Cleaved mRNA
Phosphorimage
IC$_{50}$ = 0.8 μM
R^2 = 0.97
Normalized activity (%)
120
100
80
60
40
20
0
0
20
40
60
80
100
120
Cmp2 (μM)
B
Compound 2
DMSO
HeLa NE
+
3' processing complex
H complex
Phosphorimage
C
Compound 2
DMSO
HeLa NE
-
-
+
Pre-mRNA
Cleaved mRNA
Phosphorimage
IC$_{50}$ = 100.2 μM
R^2 = 0.73
Normalized activity (%)
120
100
80
60
40
20
0
0
20
40
60
80
100
120
Cmp2 (μM)
D
Pearson r = 0.38
n = 38
IC$_{50}$ (μM)
100
80
60
40
20
0
0
10
20
30
40
50
60
PAS Activity

Figure. 3.1 Compound 2-mediated inhibition of mRNA 3’ processing in vitro is sequence-dependent

(A) In vitro cleavage assay on L3 PAS with increasing concentration of Compound 2 and its IC50 quantification. Radio-labeled RNAs from the reactions were extracted and resolved on 8M urea gel and visualized by phosphorimaging. Compound 2 concentrations used are: 0.1, 0.5, 2.5, 12.5, 62.5, and 100 μM.

(B) Electrophoretic mobility shift assay (EMSA) with L3 PAS in the presence of increasing concentration of Compound 2. Same concentrations as (A) were used.

(C) In vitro cleavage assay on SVL PAS with increasing concentration of Compound 2 and its IC50 quantification.

(D) PAS activity and IC50 correlation of 34 in vitro tested PAS.

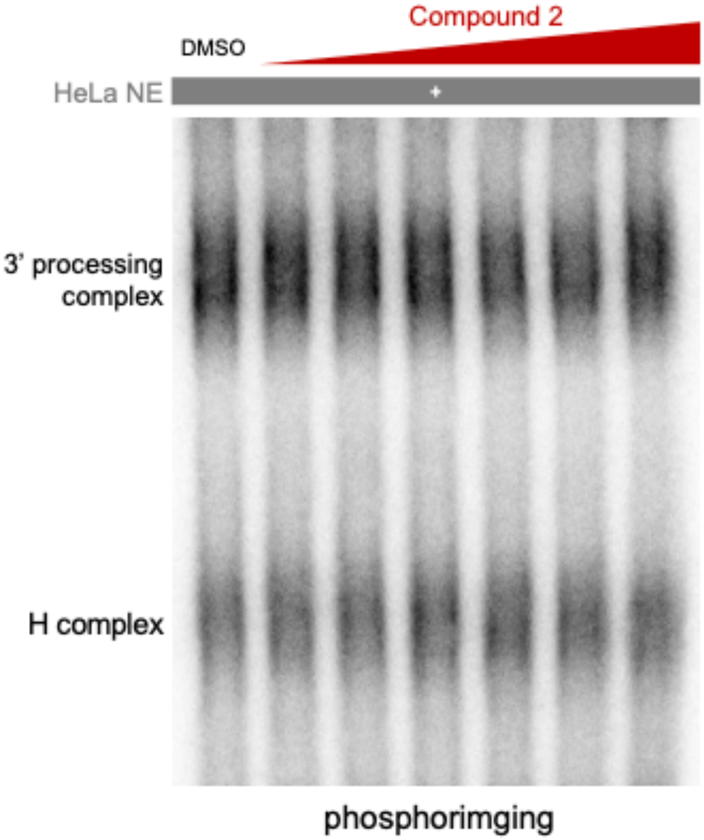

Figure 3.2 Compound 2 does not affect 3' processing complex assembly on resistant RNA.

Electrophoretic mobility shift assay (EMSA) with SVL PAS in the presence of increasing concentration of Compound 2. Same concentrations as Figure 3.1A and 3.1B were used.

The cleavage site (CS) region sequence is a major determinant of Compound 2 sensitivity

Since PASs display sequence-dependent sensitivity to Compound 2 in vitro, we next wanted to map the specific region(s) of the PAS that determine its drug sensitivity. To this end, we divided the PAS sequence into three regions: the AAUAAA hexamer and upstream sequence (referred to as upstream sequence or UPS), the CS region (20 nucleotide (nt) region centered at the cleavage site), and the downstream sequence (DS) (Figure 3.3A). Among PAS sequences we tested previously, L3 (IC_{50}=0.8 μM, Figure 3.1A) and SVL (IC_{50}=100.2 μM, Figure 3.1C) showed the lowest and the highest resistance to Compound 2, respectively. Therefore, we constructed a series of chimeric PASs between these sequences, in which one or more of the three regions in one PAS was replaced by their counterparts in another. We then measured their IC_{50} using in vitro cleavage assay as described above. Replacing the UPS of L3 PAS with that of SVL did not result in a major change in IC_{50} (Chimera 1, IC_{50}=2.1 μM, Figure 3.3A and Figure 3.4A). However, replacing both the UPS and the CS of L3 with those of SVL dramatically increased the resistance to Compound 2 (Chimera 2, IC_{50}=89.6 μM, Figure 3.3A and Figure 3.4B), suggesting that the CS region plays a major role. On the other hand, replacing the UPS of SVL with that of L3 led to a significant decrease in drug resistance (Chimera 3, Figure 3.3A and Figure 3.4C), although its IC_{50} (39.5 μM) was still nearly 50 times higher than that of L3. Replacing both the UPS and CS of SVL with those of L3 led to a near 15-fold decrease in IC_{50} (Chimera 4, IC_{50}=6.7 μM, Figure 3.3A and Figure 3.4D), again highlighting a major role for the CS region. By contrast, the DS did not seem to play a significant role (compare L3 and Chimera 4 or SVL and Chimera 2, Figure 3.3A). Given the large impact of the CS region on Compound 2 sensitivity in both backgrounds, we swapped the CS regions alone between L3 and SVL. The results showed that replacing the L3 CS region with that of SVL increased its IC_{50} to 47.8 μM, a

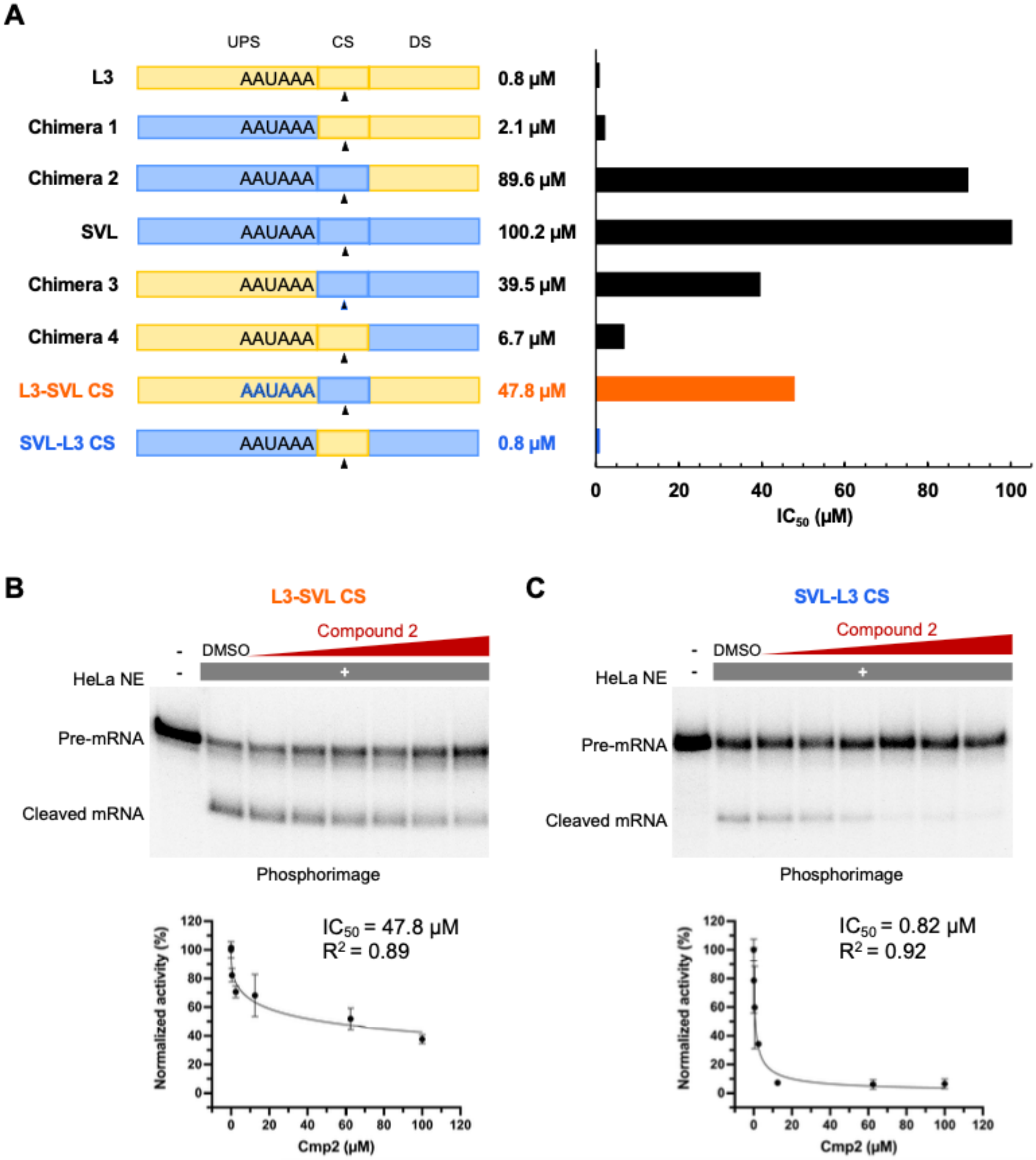
A
UPS
CS
DS
L3
Chimera 1
Chimera 2
SVL
Chimera 3
Chimera 4
L3-SVL CS
SVL-L3 CS
AAUAAA
0.8 μM
2.1 μM
89.6 μM
100.2 μM
39.5 μM
6.7 μM
47.8 μM
0.8 μM
0
20
40
60
80
100
IC50 (μM)
B
L3-SVL CS
Compound 2
-
DMSO
HeLa NE
-
+
Pre-mRNA
Cleaved mRNA
Phosphorimage
IC50 = 47.8 μM
R2 = 0.89
Normalized activity (%)
Cmp2 (μM)
C
SVL-L3 CS
Compound 2
-
DMSO
HeLa NE
-
+
Pre-mRNA
Cleaved mRNA
Phosphorimage
IC50 = 0.82 μM
R2 = 0.92
Normalized activity (%)
Cmp2 (μM)

Figure 3.3. Cleavage site (CS) region is a major determinant of Compound 2 sensitivity

(A) A diagram of L3, SVL and their chimeras. Their corresponding IC_{50} were plotted on the right. UPS: upstream sequence; CS: cleavage site; DS: downstream sequence. Black triangles denote the cleavage position YA (Y is U or C).

(B) In vitro cleavage of L3-SVL CS with increasing concentration of Compound 2, similar to Figure 3.1A and C.

(C) In vitro cleavage of SVL-L3 CS with increasing concentration of Compound 2.

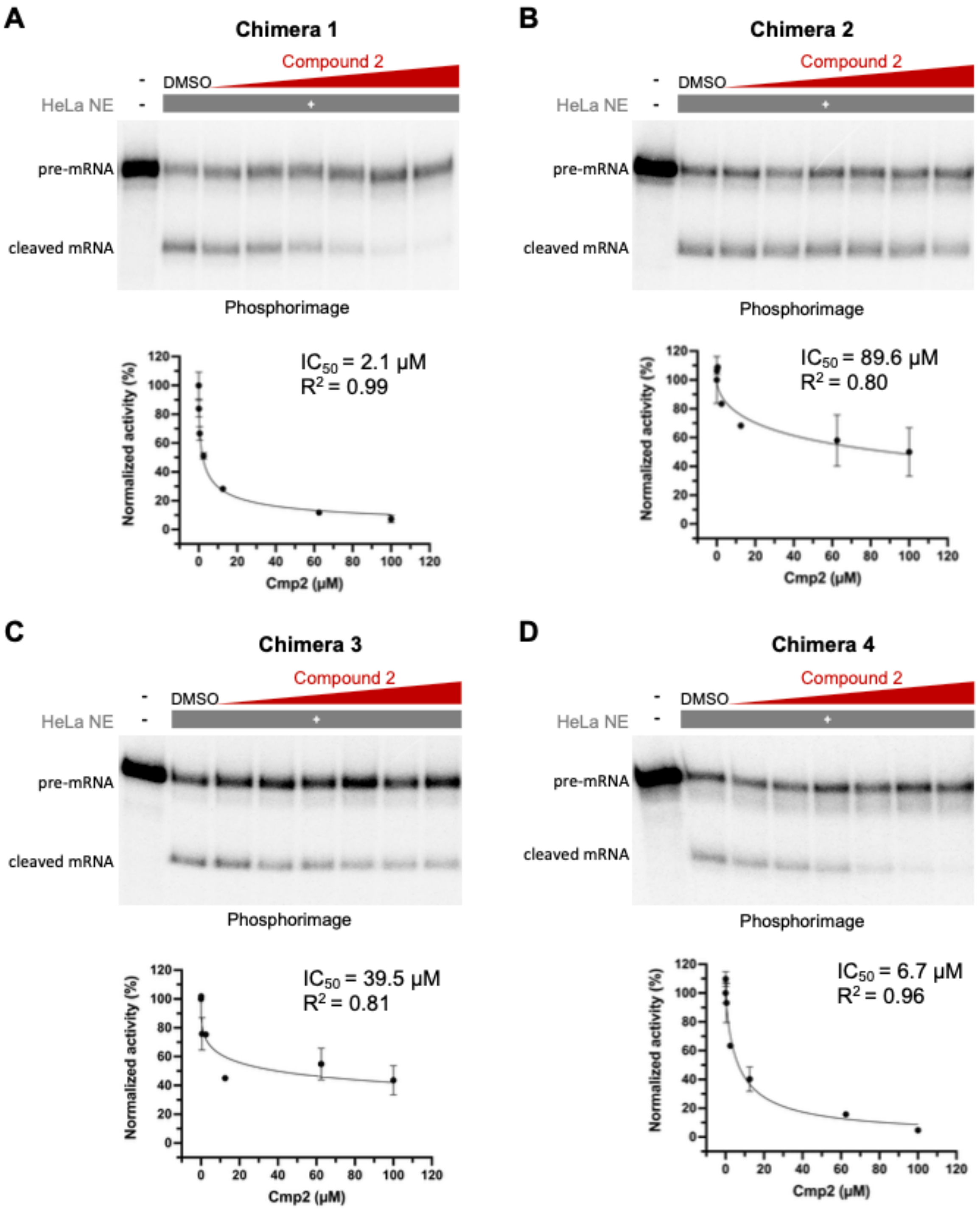

Figure 3.4. In vitro cleavage for L3 and SVL chimeras

(A)~(D). In vitro cleavage of L3-SVL chimeras 1~4 with increasing concentration of Compound 2 and their IC50, similar to Figure 3.1 and 3.3.

near 60-fold increase (L3-SVL CS, Figure 3.3A and B). Even more dramatically, the opposite change in SVL reduced its IC_{50} to 0.8 μM (SVL-L3 CS, Figure 3.3A and C), identical to that of L3. These results demonstrated that the CS region is a major determinant of Compound 2 sensitivity in both backgrounds. Additionally, the UPS also contributes to the drug sensitivity in a context-dependent manner, while the DS does not appear to play a significant role. Therefore, we have focused on the CS region for the rest of this study.

Define the CS sequence-Compound 2 sensitivity relationship by using massively parallel in vitro assay (MPIVA)

We next wanted to comprehensively define the relationship between the CS region sequence and Compound 2 sensitivity. To this end, we designed an MPIVA strategy (Figure 3.5A). Using L3 (sensitive) or SVL (resistant) PAS as backbones, we replaced the original cleavage site with a YA sequence (Y is U or C), which is the preferred cleavage site for CPSF73, and randomized the 23nt flanking sequence. We have found that just changing the cleavage site from UA to CA (for L3) and CA to UA (for SVL) does not affect the sensitivity of PAS to Compound 2 (Figure 3.6). Changing the cleavage site also did not alter the cleavage efficiency for L3 but reduced SVL cleavage efficiency by 14% (data not shown)**.** These two libraries, called L3-N23 and SVL-N23, contained ~3 million PAS variants each and were transcribed into RNA. The PAS RNA pools were used for in vitro cleavage and polyadenylation assays in the presence of DMSO (control) or increasing concentrations of Compound 2, including low (0.5 μM) , medium (2.5 μM), and high (12.5 μM) concentrations. As shown in Figure 3.5B, the PAS RNA pool was efficiently cleaved in vitro in the presence of DMSO and the cleavage efficiency gradually decreased in the presence of increasing concentrations of Compound 2. The starting

PAS RNA pool and the cleaved RNA pools under different conditions were subjected to high throughput sequencing using the Illumina platform (Figure 3.5B). For each variant, a resistance score was calculated as the log ratio between its frequency in Compound 2-treated samples and that in DMSO-treated samples. As shown in Figure 3.5C and Figure 3.7A, the resistance scores of all variants were concentrated in a narrow peak centered at ~0 at low Compound 2 concentration (L3: -0.04 ± 0.31; SVL: -0.05 ± 0.28) but diverged more at high inhibitor concentration (L3: -0.12 ± 0.53; SVL: -0.09 ± 0.43), suggesting that, as expected, drug sensitivities are better distinguished at higher drug concentrations. Furthermore, we compared the resistance scores of all variants and their cleavage efficiency (log ratio between the frequency of a PAS variant in Library 2 and that in Library 1) and found that there was no significant correlation (Figure 3.5D and Figure 3.7B), which was consistent with Figure 3.1D. Thus, both our low throughput in vitro assays and high throughput screen results demonstrated that the Compound 2 sensitivity of a PAS is not dependent on its strength.

Based on the resistance scores in the high Compound 2 concentration condition, we obtained a list of the top 1000 most sensitive and resistant PASs from both the L3-N23 and SVL-N23 libraries. We selected 6 variants, 3 sensitive and 3 resistant, in each background and tested them using our in vitro cleavage assay and our data validated the screen results (Figure 3.5E, Figure 3.8). It was noted that some of the variants (e.g. Figure 3.5E, top left panel) were more sensitive to Compound 2 than the original L3 while other variants displayed greater resistance than SVL (e.g. Figure 3.5E, bottom right panel), indicating that our screens selected variants with a wide range of drug sensitivities. Interestingly, the nucleotide composition in the CS region of sensitive and resistant PASs showed distinct patterns. The CS regions of sensitive L3 variants are generally G/U-rich, especially in the region upstream of the cleavage site (Figure 3.5F, top panel). By contrast, resistant CSs contained alternating A-rich and U-rich sequences mainly in the region upstream of the cleavage site (Figure 3.5F, bottom panel). Very similar patterns were observed in SVL background (Figure 3.5G), suggesting that the CS region sequence can determine Compound 2 sensitivity independent of other regions. Consistent with the nucleotide compositions, our motif analyses of the sensitive and resistant variants detected U/G-rich and A/U-rich motifs, respectively, in both L3 and SVL libraries (Figure 3.9). These results defined the key sequence features in the CS region that determine Compound 2 sensitivity.

Machine learning predictions of Compound 2 sensitivity from PAS sequences

We next used our MPIVA data to train a machine learning model with the goal of predicting Compound 2 sensitivity of any given PAS based on its CS region sequence. Our model, called **C**leavage and **C**ounteraction with **C**ompound 2 on **P**olyadenylation **O**utcomes (C3PO), is a three-layer convolutional neural network (CNN) that is based on the Optimus 5′ architecture that we have previously used to predict polysome profiles from 5′ untranslated region (UTR) sequences (Figure 3.10A, methods) (Sample et al. 2019). C3PO uses the 25 nt CS sequences as inputs and predicts Compound 2 sensitivity, which is calculated as the log ratio between each variant's percent representation in the DMSO-treated and Compound 2-treated libraries (Figure 3.5A). C3PO was trained on the processed MPIVA datasets from both the L3 and SVL RNA contexts, and model performance was assessed on held-out variants from both RNA contexts. We used the variants with high read coverage in the input and DMSO-treated data (Libraries 1 and 2) as our test set to minimize the impact of measurement noise (methods). C3PO performed better on higher doses of Compound 2 with Pearson's r of 0.56, 0.74, and 0.84 for 0.5 μM, 2.5 μM, and 12.5 μM, respectively. We explored variations of convolution-based machine learning

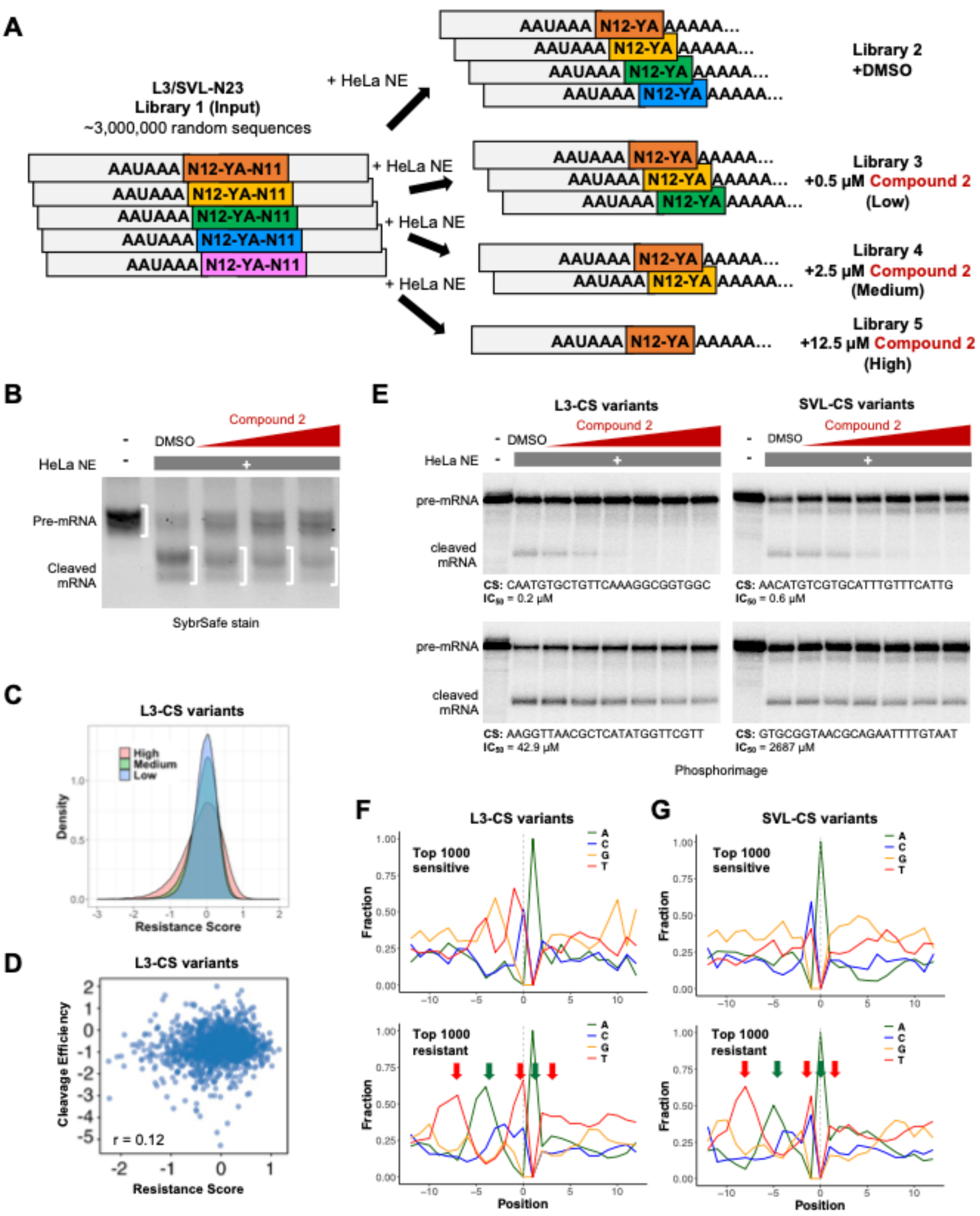
A
L3/SVL-N23
Library 1 (Input)
~3,000,000 random sequences
AAUAAA N12-YA-N11
AAUAAA N12-YA-N11
AAUAAA N12-YA-N11
AAUAAA N12-YA-N11
AAUAAA N12-YA-N11
+ HeLa NE
+ HeLa NE
+ HeLa NE
+ HeLa NE
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
Library 2
+DMSO
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
Library 3
+0.5 µM Compound 2
(Low)
AAUAAA N12-YA AAAAA…
AAUAAA N12-YA AAAAA…
Library 4
+2.5 µM Compound 2
(Medium)
AAUAAA N12-YA AAAAA…
Library 5
+12.5 µM Compound 2
(High)
B
Compound 2
- DMSO
HeLa NE - +
Pre-mRNA
Cleaved mRNA
SybrSafe stain
C
L3-CS variants
High
Medium
Low
Density
1.0
0.5
0.0
-3 -2 -1 0 1 2
Resistance Score
D
L3-CS variants
Cleavage Efficiency
2 1 0 -1 -2 -3 -4 -5
r = 0.12
-2 -1 0 1
Resistance Score
E
L3-CS variants
SVL-CS variants
Compound 2
- DMSO
HeLa NE - +
pre-mRNA
cleaved mRNA
CS: CAATGTGCTGTTCAAAGGCGGTGGC
IC$_{50}$ = 0.2 µM
CS: AACATGTCGTGCATTTGTTTCATTG
IC$_{50}$ = 0.6 µM
pre-mRNA
cleaved mRNA
CS: AAGGTTAACGCTCATATGGTTCGTT
IC$_{50}$ = 42.9 µM
CS: GTGCGGTAACGCAGAATTTTGTAAT
IC$_{50}$ = 2687 µM
Phosphorimage
F
L3-CS variants
G
SVL-CS variants
Top 1000 sensitive
Top 1000 resistant
A
C
G
T
Fraction
1.00
0.75
0.50
0.25
0.00
-10 -5 0 5 10
Position

Figure 3.5 Determine sequence specificity for Compound 2 sensitivity by massively parallel in vitro assay (MPIVA)

(A) Design of the randomized CS sequence libraries and the MPIVA assay. Each box represents a sequence variant. YA: cleavage position (Y is U or C). N: random nucleotide.

(B) The randomized sequence library L3/SVL-N23 were transcribed into RNAs and used for in vitro cleavage/polyadenylation assays in the presence of 0.5, 2.5, and 12.5 μM Compound 2. The RNAs from these reactions were amplified by RT-PCR and resolved on an agarose gel. The RNA species were marked on the left. The white half brackets mark the regions on the gel that were extracted and amplified for sequencing.

(C) A density plot for the resistance scores of all variants in L3-N23 library. The low, medium, and high groups represent the screens in the presence of 0.5, 2.5, and 12.5 μM Compound 2 as shown in (B).

(D) A scatter plot comparing the cleavage efficiency log(frequency in Library 2/frequency in Library 1) and the resistance score (log(frequency in Library 5/frequency in Library 2) of L3-CS variants. Pearson correlation is shown.

(E) Examples of validation experiments using in vitro cleavage assays for variants from both L3- and SVL-N23 libraries.

(F) Nucleoside distribution of L3-CS variants for the top 1000 most sensitive and resistant sequences.

(G) Nucleoside distribution of SVL-CS variants for the top 1000 most sensitive and resistant sequences. T- and A- rich regions were marked with red and green arrows respectively.

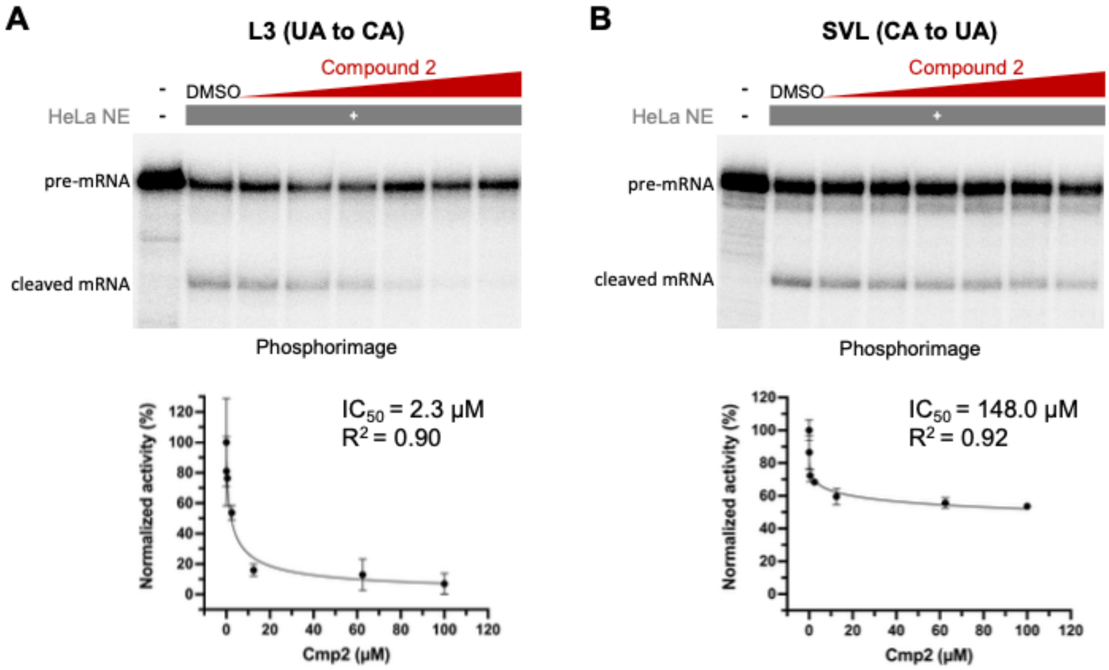

Figure 3.6 Mutating the YA cleavage site does not affect PAS sensitivity to Compound 2

In vitro cleavage on **(A).** L3 (UA to CA mutant) and **(B).** SVL (CA to UA mutant) and their IC$_{50}$. Compound 2 concentration used is the same as Figure 3.1 and 3.3.

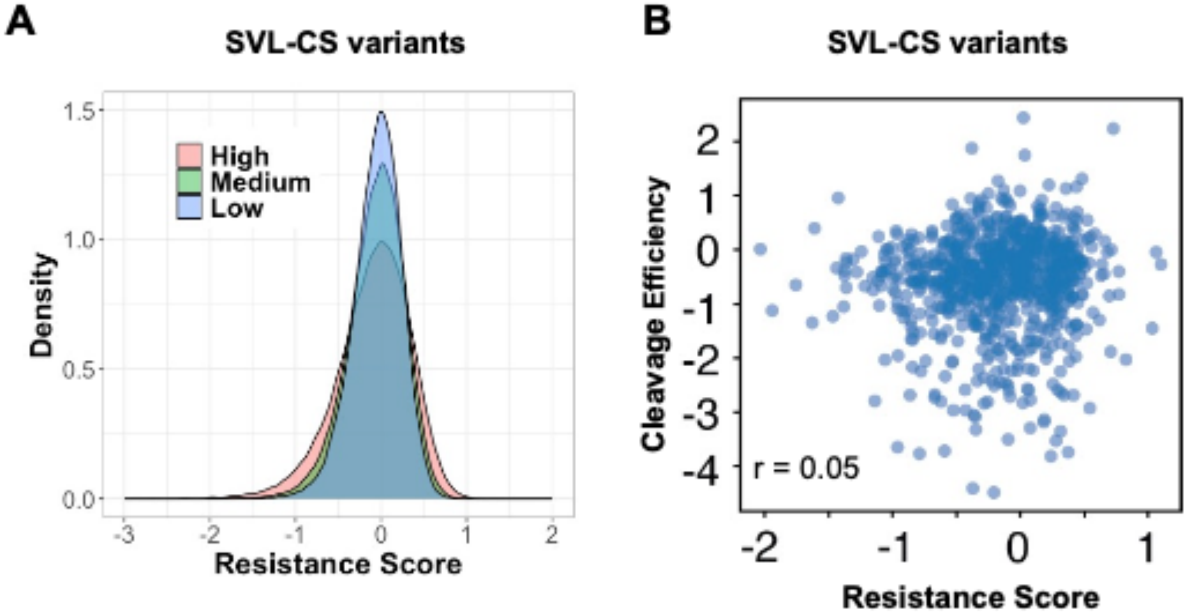

Figure 3.7. Characterization of SVL variants by MPIVA

(A) A density plot for the resistance scores of all variants in SVL-N23 library. The low, medium, and high groups represent the screens in the presence of 0.5, 2.5, and 12.5 μM Compound 2.

(B) A scatter plot comparing the cleavage efficiency log(frequency in Library 2/frequency in Library 1) and the resistance score (log(frequency in Library 5/frequency in Library 2) of SVL-CS variants. Pearson correlation is shown.

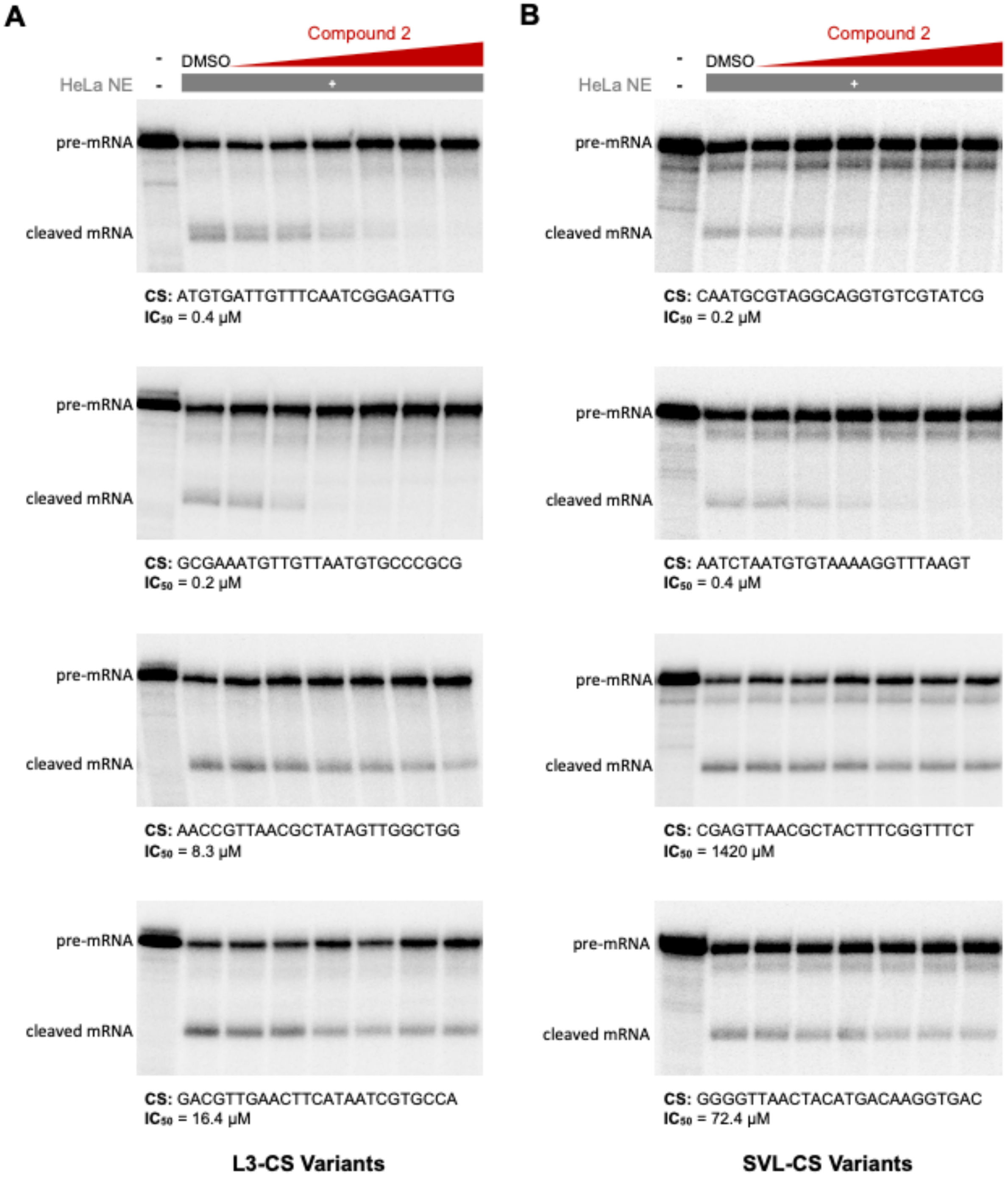
A
Compound 2
-
DMSO
HeLa NE
-
+
pre-mRNA
cleaved mRNA
CS: ATGTGATTGTTTCAATCGGAGATTG
IC₅₀ = 0.4 µM
pre-mRNA
cleaved mRNA
CS: GCGAAATGTTGTTAATGTGCCCGCG
IC₅₀ = 0.2 µM
pre-mRNA
cleaved mRNA
CS: AACCGTTAACGCTATAGTTGGCTGG
IC₅₀ = 8.3 µM
pre-mRNA
cleaved mRNA
CS: GACGTTGAACTTCATAATCGTGCCA
IC₅₀ = 16.4 µM
L3-CS Variants
B
Compound 2
-
DMSO
HeLa NE
-
+
pre-mRNA
cleaved mRNA
CS: CAATGCGTAGGCAGGTGTCGTATCG
IC₅₀ = 0.2 µM
pre-mRNA
cleaved mRNA
CS: AATCTAATGTGTAAAAGGTTTAAGT
IC₅₀ = 0.4 µM
pre-mRNA
cleaved mRNA
CS: CGAGTTAACGCTACTTTCGGTTTCT
IC₅₀ = 1420 µM
pre-mRNA
cleaved mRNA
CS: GGGGTTAACTACATGACAAGGTGAC
IC₅₀ = 72.4 µM
SVL-CS Variants

Figure 3.8 Additional validated CS variants from both backbones from in vitro MPIVA

In vitro cleavage validation experiment of 4 more RNA (2 sensitive and 2 resistant) from **(A).** L3- and **(B).** SVL-N23 libraries. Compound 2 concentration used is the same as Figure 3.1 and 3.3. The CS region sequence and their IC_{50} is shown.

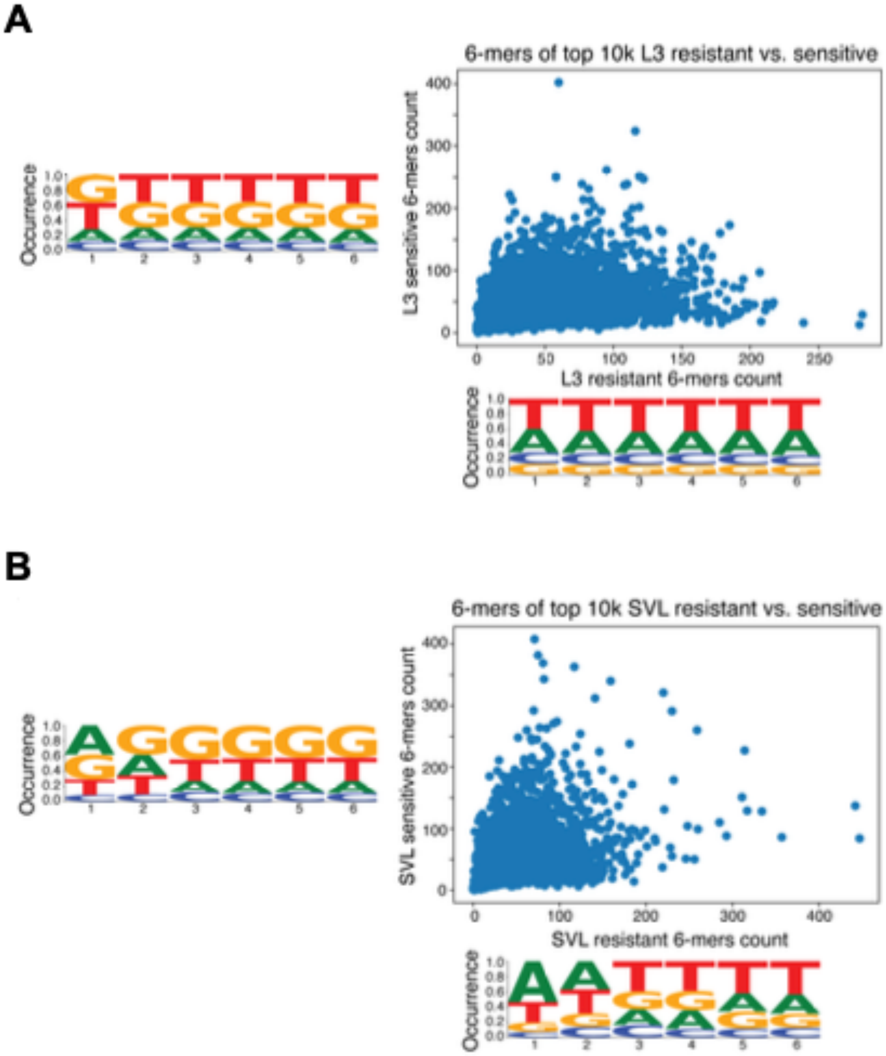

Figure 3.9 6-mer motif analyses of top 10,000 resistant and sensitive CS variants from both backbones from the in vitro MPIVA

Counts of 6-mers from **(A)** L3 and **(B)** SVL backbones are plotted alongside the nucleotide content of significantly enriched 6-mers in the top sensitive (left logo) and resistant (bottom logo) 10,000 CS variants. Sequence logos use DNA-encoding of RNA nucleotides.

architectures (Table 3.2), and this trend was consistent. This was expected as drug resistance is better detected at higher drug dose (Figure 3.5C). Due to the better model performance at the highest dose of 12.5 µM, we focused further analyses on this regime.

To test the performance of C3PO, we compared the Compound 2 resistance scores (log(12.5 µM/DMSO)) of 30 distinct PASs measured by in vitro cleavage assays as shown in Figure 3.1D (PASs that contain the same CS region sequences were omitted to avoid redundancy) and those predicted by C3PO. The C3PO predictions showed strong and positive correlation with experimental measurements with a Pearson r of 0.84 (Figure 3.10C, Table 3.3). This is very similar to its performance on the MPIVA dataset (compare Figure 3.10C with 3.10B, 12.5 µM panel). These results strongly suggest that C3PO can accurately predict Compound 2 sensitivity of PAS sequence in vitro.

We next wanted to identify sequence motifs that are predictive of Compound 2 sensitivity by extracting filter position weight matrices (Figure 3.10A). The position-specific effect on Compound 2 sensitivity of each filter was quantified by measuring the correlation with drug sensitivity at each position across the CS region. Filters associated with higher resistance (dark red color) learned motifs that were A/U-rich, while lower resistance filters (dark blue) typically learned motifs with higher G/U content (Figure 3.10D). Sequence motifs strongly associated with Compound 2 sensitivity predictions are positioned such that they begin upstream of the CS (Figure 3.10D-F, Figure 3.11). Layer 2 filters learn to use combinations of Layer 1 filters for predictions of drug sensitivity. 15-mers learned by layer 2 filters also showed A/U-rich and G/U-rich motifs for resistant and sensitive PASs respectively (Figure 3.10E and Figure 3.12). Interestingly, both resistance- and sensitivity-associated motifs are enriched in the region upstream of the CS (Figure 3.10F).

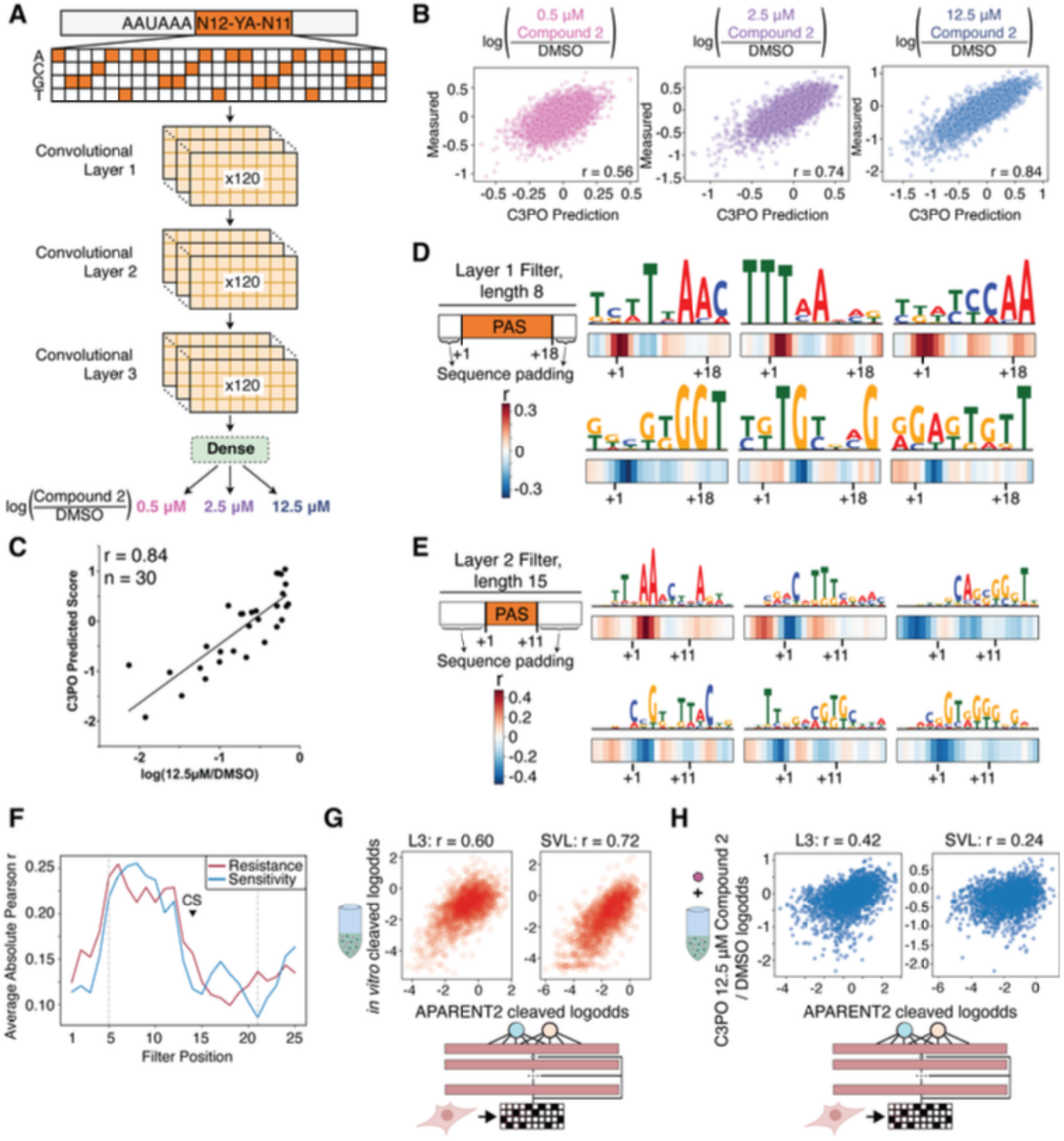
A
AAUAAA N12-YA-N11
A
C
G
T
Convolutional Layer 1
x120
Convolutional Layer 2
x120
Convolutional Layer 3
x120
Dense
log(Compound 2 / DMSO) 0.5 µM 2.5 µM 12.5 µM
B
log(0.5 µM Compound 2 / DMSO)
log(2.5 µM Compound 2 / DMSO)
log(12.5 µM Compound 2 / DMSO)
Measured
C3PO Prediction
r = 0.56
r = 0.74
r = 0.84
C
r = 0.84
n = 30
C3PO Predicted Score
log(12.5µM/DMSO)
D
Layer 1 Filter, length 8
PAS
+1
+18
Sequence padding
r
0.3
0
-0.3
E
Layer 2 Filter, length 15
PAS
+1
+11
Sequence padding
r
0.4
0.2
0
-0.2
-0.4
F
Average Absolute Pearson r
Resistance
Sensitivity
CS
Filter Position
G
L3: r = 0.60
SVL: r = 0.72
in vitro cleaved logodds
APARENT2 cleaved logodds
H
L3: r = 0.42
SVL: r = 0.24
C3PO 12.5 µM Compound 2 / DMSO logodds
APARENT2 cleaved logodds

Figure 3.10 C3PO architecture, performance, and layer feature analyses

(A) The model takes 25 nt RNA sequences immediately downstream of the core hexamer and predicts three doses of Compound 2 drug sensitivity by predicting the log ratio of percent reads in a drug-treated sample to a DMSO-treated sample.

(B) Scatter plots of C3PO performance on predicting drug sensitivity at 3 Compound 2 doses on test sequences. Test sequences include equal number of sequences derived from both the L3 and SVL RNA contexts.

(C) A scatter plot comparing the resistance scores predicted by C3PO and those measured experimentally.

(D) Convolutional layer 1 and **(E)** layer 2 max filter activations with the highest Pearson correlation with 12.5 µM Compound 2 predictions. Sequence logos are plotted on top of per-position absolute value of Pearson correlations with 12.5 µM Compound 2 sensitivity predictions. All layer 1 and 2 filters are reported in Figure 3.11 and 3.12

(F) Plot of average of all layer 1 filters' absolute value of Pearson correlation with 12.5 µM Compound 2 predictions across all positions. These are split into Pearson correlation values associated with resistance or sensitivity. Dashed gray lines indicate positions at the edge of sequence padding. The position of the cleavage site (CS) is marked and note that preceding filters may overlap with the designed canonical cut sites.

(G) Scatterplots of RNA cleavage logodds measured in vitro calculated from input and DMSO libraries versus those from APARENT2 predictions.

(H) Scatterplots of Compound 2 resistance predicted by C3PO and the cleavage efficiency predicted by APARENT2.

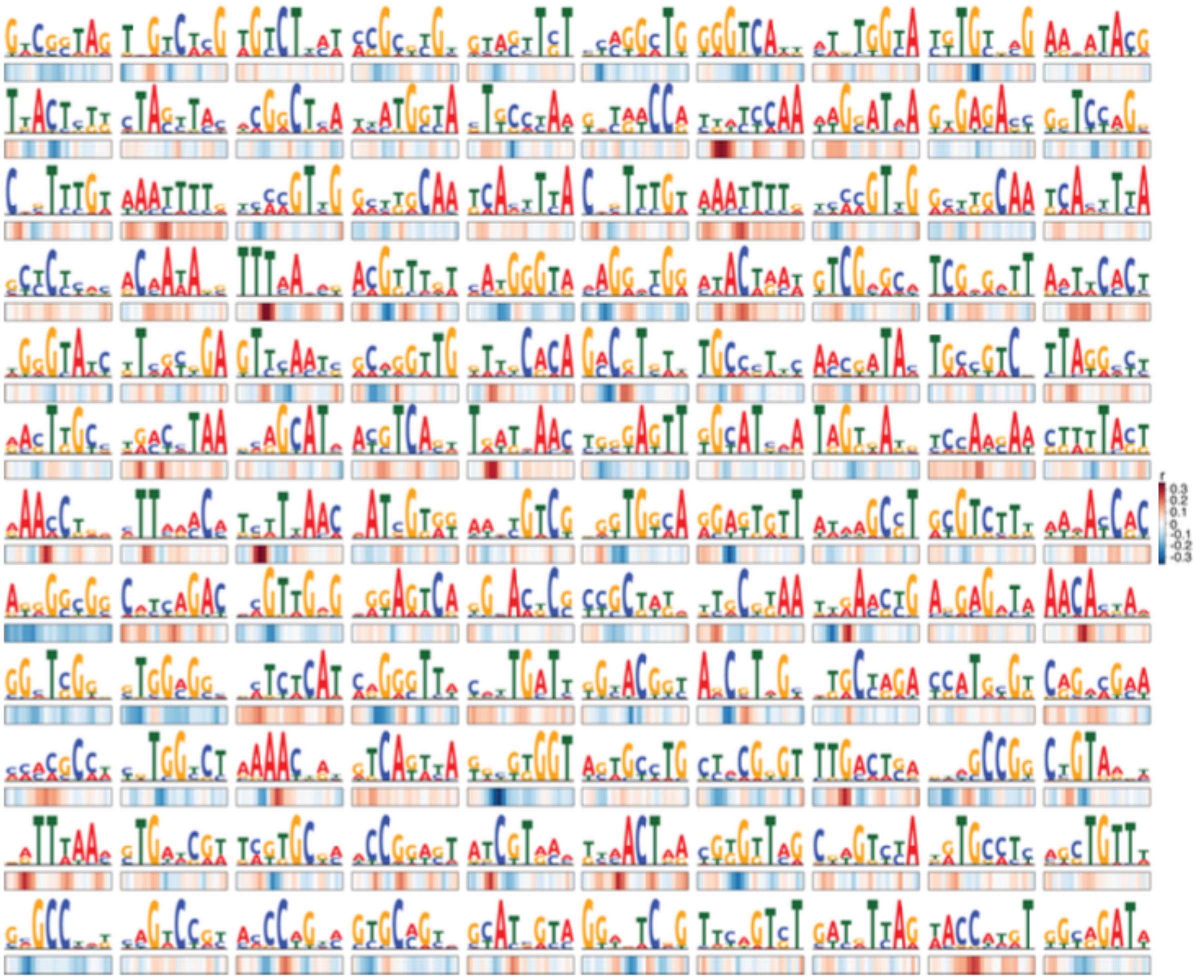

Figure 3.11 C3PO layer 1 learned sequence features

All of C3PO’s layer 1 filters’ max activation sequence consensus and correlations with 12.5 µM Compound 2 sensitivity predictions. Related to Figure 3.10D.

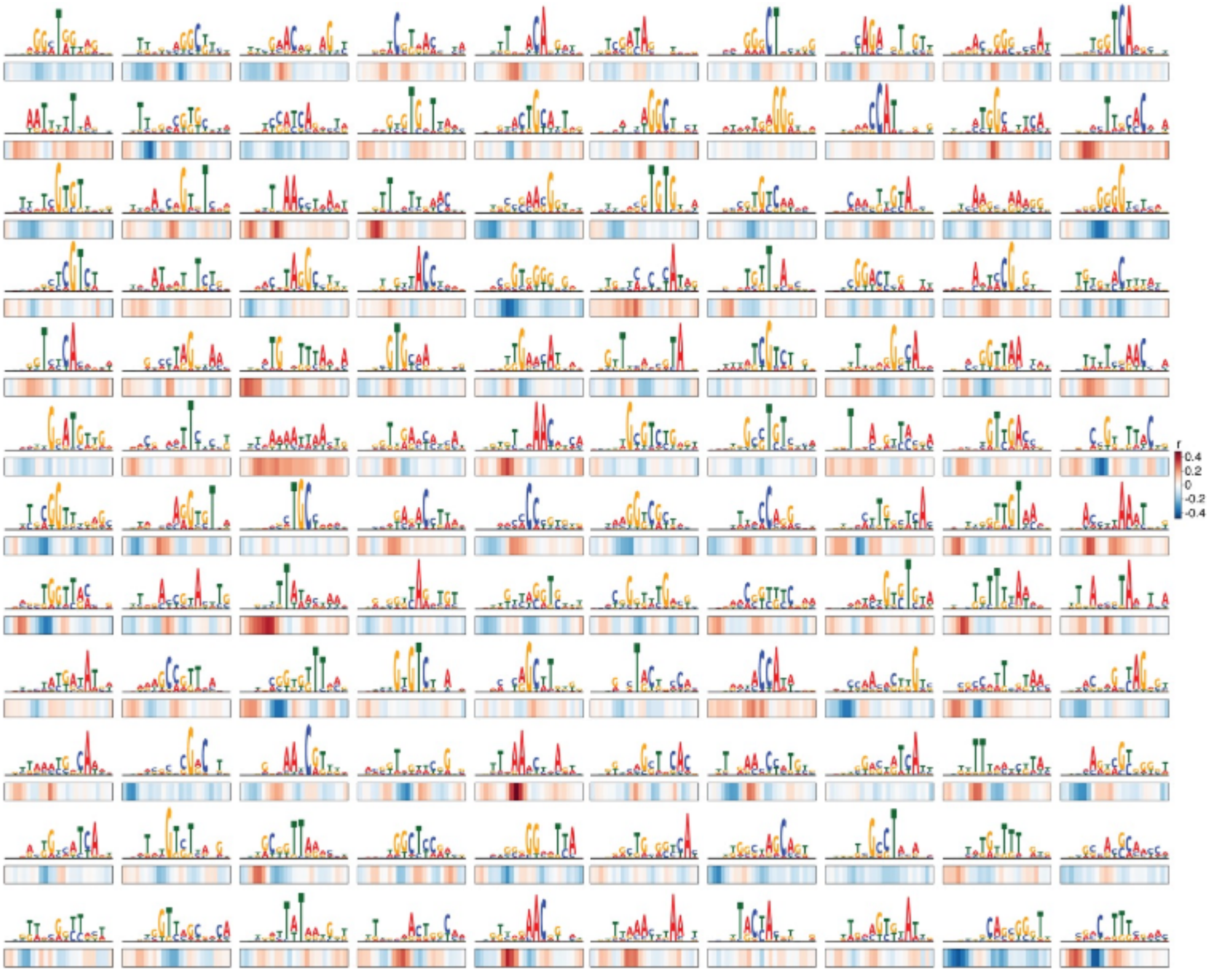

Figure 3.12 C3PO layer 2 learned sequence features

All of C3PO’s layer 2 filters’ max activation sequence consensus and correlations with 12.5 μM Compound 2 sensitivity predictions. Related to Figure 3.10E.

Given the known function of RNA secondary structures in pre-mRNA 3′ processing (Wu and Bartel 2017), we investigated the its potential impact on Compound 2 sensitivity. We compared the minimum free energy (MFE) structures for the top 10,000 resistant and sensitive sequences (Figure 3.13A-B). The differences between ΔGs for the resistant and sensitive sequences were modest, but statistically significant with p-values of $< 2.2 \times 10^{-308}$ and 1.58×10^{-26} and for L3 and SVL, respectively. The difference between base pairing probabilities for resistant and sensitive sequences also show different global patterns between the L3 and SVL backbones, indicating that background-specific secondary structural features may contribute to drug sensitivity (Figure 3.13C-D). Taken together with C3PO's ability to accurately predict Compound 2 sensitivity with sequence alone, our results suggest that sequence is the primary determinant of Compound 2 sensitivity while secondary structure may play a minor role.

We further explored the usage of machine learning models to characterize Compound 2 sensitivity and its relationship with processing efficiency. First, we compared the cleavage efficiency measured by our MPIVA assays with that predicted by APARENT2 (Linder et al. 2022), a highly accurate deep learning model for predicting cleavage/polyadenylation efficiency that was trained using massively parallel reporter assays in mammalian cells. We saw good correlation between APARENT2-predicted cleavage efficiency and our MPIVA data with a Pearson r of 0.60 for the L3 background and 0.72 for the SVL background (Figure 3.10G). These results suggest that the CS region sequence can have a significant impact on cleavage efficiency, and that the cleavage efficiency values measured by our MPIVA system are highly consistent with measurements obtained in cells. Finally, we compared the resistance score predicted by C3PO with the cleavage efficiency predicted by APARENT2 for all CS variants and observed poor correlation with Pearson r = 0.42 and 0.24 for L3 and SVL respectively (Figure 3.10H).

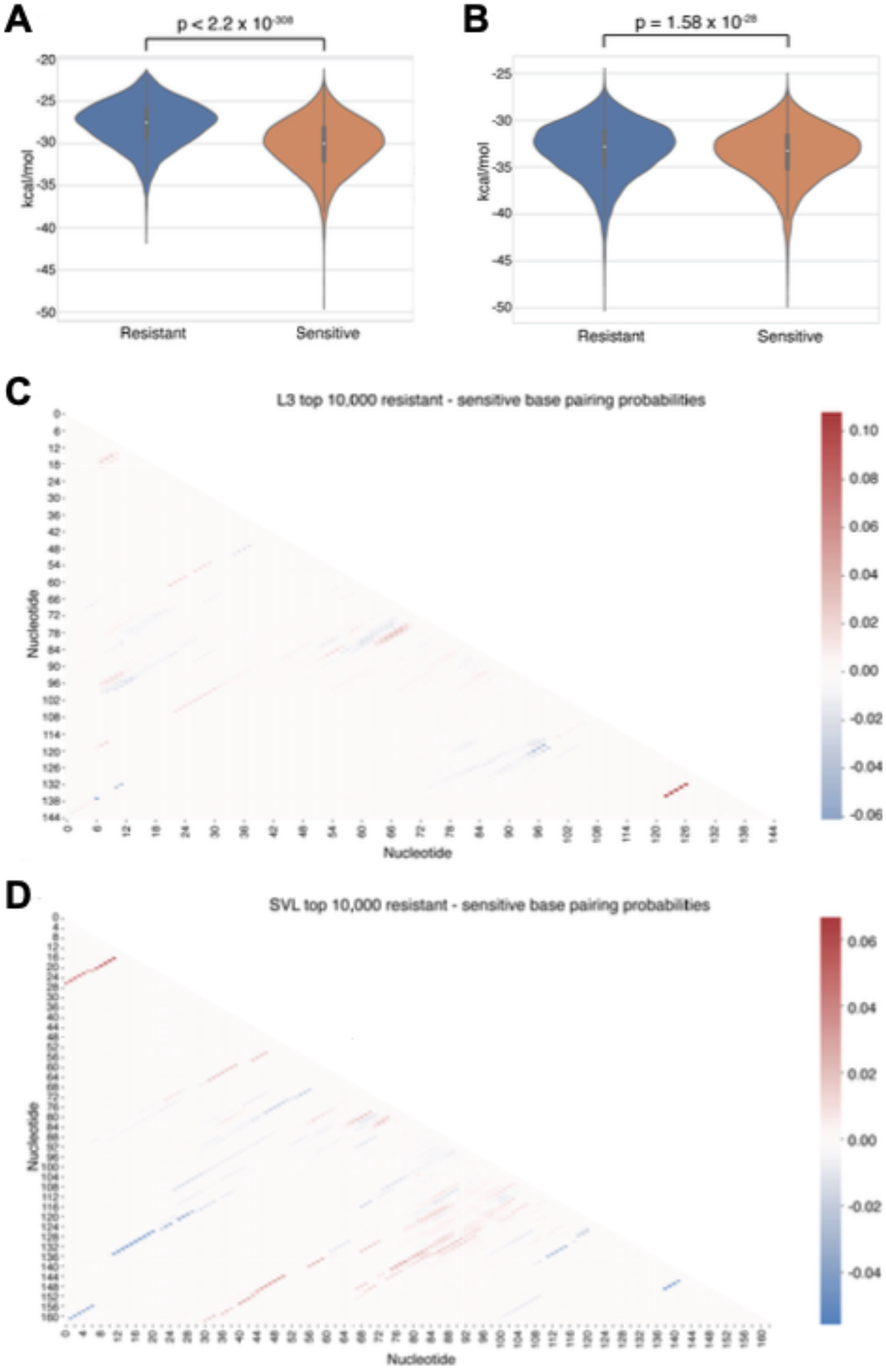
A
p < 2.2 x 10-308
kcal/mol
-20
-25
-30
-35
-40
-45
-50
Resistant
Sensitive
B
p = 1.58 x 10-28
kcal/mol
-25
-30
-35
-40
-45
-50
Resistant
Sensitive
C
L3 top 10,000 resistant - sensitive base pairing probabilities
Nucleotide
Nucleotide
0.10
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
D
SVL top 10,000 resistant - sensitive base pairing probabilities
Nucleotide
Nucleotide
0.06
0.04
0.02
0.00
-0.02
-0.04

Figure 3.13. DG of minimum free energy structures and base pairing probabilities of the top 10,000 resistant and sensitive sequences

Comparison of minimum free energy (MFE) structures' of DG's from the top 10,000 resistant and sensitive **(A)** L3 and **(B)** SVL sequences. The DG's are significant with a p- value of < 2.2 x 10^{-308} for L3 and 1.58 x 10^{-28} for SVL (t-test with unequal variance).

(C) Heatmap of the difference between top 10,000 resistant and sensitive L3 sequences' average base pairing probabilities.

(D) Same as in panel C but for SVL.

This is consistent with our in vitro cleavage assay (Figure 3.1D) and MPIVA results (Figure 3.5D and Figure 3.7B) and provided further evidence that the Compound 2 sensitivity of a PAS is not dependent on its strength.

JTE-607 modulates PAS selection and transcription termination in a sequence-specific manner in human cells

To determine whether the sequence-specific sensitivity to Compound 2 observed in vitro was true in cells, we performed two genome-wide analyses. First, we analyzed the global APA profiles in DMSO- and JTE-607-treated human HepG2 cells using PAS-seq, a high throughput RNA 3′ sequencing method for quantitatively mapping RNA polyadenylation (Yoon et al. 2021). JTE-607 treatment induced significant APA changes in 921 genes, of which 847 genes (92%) shifted from a proximal PAS to a distal one (blue dots, Figure 3.14A and see Methods for details). An example was shown in Figure 3.14B: the proximal PAS was predominantly used for *Ptp4a1* transcripts in DMSO-treated cells. However, polyadenylation shifted to a distal PAS in JTE-607 treated cells, leading to 3′ UTR lengthening. 74 genes showed APA changes in the opposite direction (red dots, Figure 3.14A), as exemplified by *Paqr8* (Figure 3.14C).

Why did JTE-607 induce the opposite APA changes in different groups of genes? Given our finding that JTE-607-mediated inhibition of mRNA 3′ processing is sequence-specific, we predicted that JTE-607 treatment would decrease the usage of the more sensitive PASs in a given gene while the usage of resistant PASs would be less impacted, leading to a net shift to the more resistant PASs. Therefore, we hypothesized that the directionality of JTE-607-induced APA change in any given gene is determined by the relative sensitivities of its alternative PASs. To test this hypothesis, we predicted the resistance scores of all annotated PASs in the human genome using C3PO and compared the scores of the proximal and distal PASs of the 921 genes that displayed significant APA shifts in JTE-607 treated cells. Interestingly, for the 847 genes that showed a shift to the distal PAS in JTE-607-treated cells, their proximal PASs are significantly more sensitive to JTE-607 than their distal ones ($p < 2.2 \times 10^{-16}$, t-test, Figure 3.14D, left panel). The opposite trend was observed for the 74 genes that showed a distal-to-proximal shift ($p = 0.03$, t-test, Figure 3.14D, right panel). Therefore JTE-607 indeed inhibited the usage of more sensitive PASs, resulting in higher usage of resistant PASs. These data confirmed that JTE-607 modulates PAS selection globally in a sequence-dependent manner in human cells and showed that JTE-607-induced APA changes depend on the relative drug sensitivities of the alternative PASs.

Additionally, we also monitored transcription termination by nascent RNA sequencing using 4-thiouridine labeled RNA (4sU-seq) in HepG2 cells treated with DMSO or JTE-607 for 4 hours and 4sU for 1 hour at 37˚C. As mRNA 3′ processing is coupled to transcription termination, transcription termination efficiency at PAS can be used as a proxy for mRNA 3' processing efficiency (Nojima et al. 2015). Our 4sU-seq analyses showed that JTE-607 treatment induced a global transcription termination defect (Figure 3.15A). However, the levels of JTE-607-induced transcription readthrough (RT) varied widely at different PASs (Figure 3.15A). For example, RT increased dramatically downstream of the PAS of the *Eif4ebp1* gene (Figure 3.15B, left panel) while little change was observed for *Cox8A* gene (Figure 3.15B, right panel). Thus 4sU-seq data further demonstrated that mRNA 3′ processing displayed sequence-specific sensitivity to JTE-607-mediated inhibition in human cells.

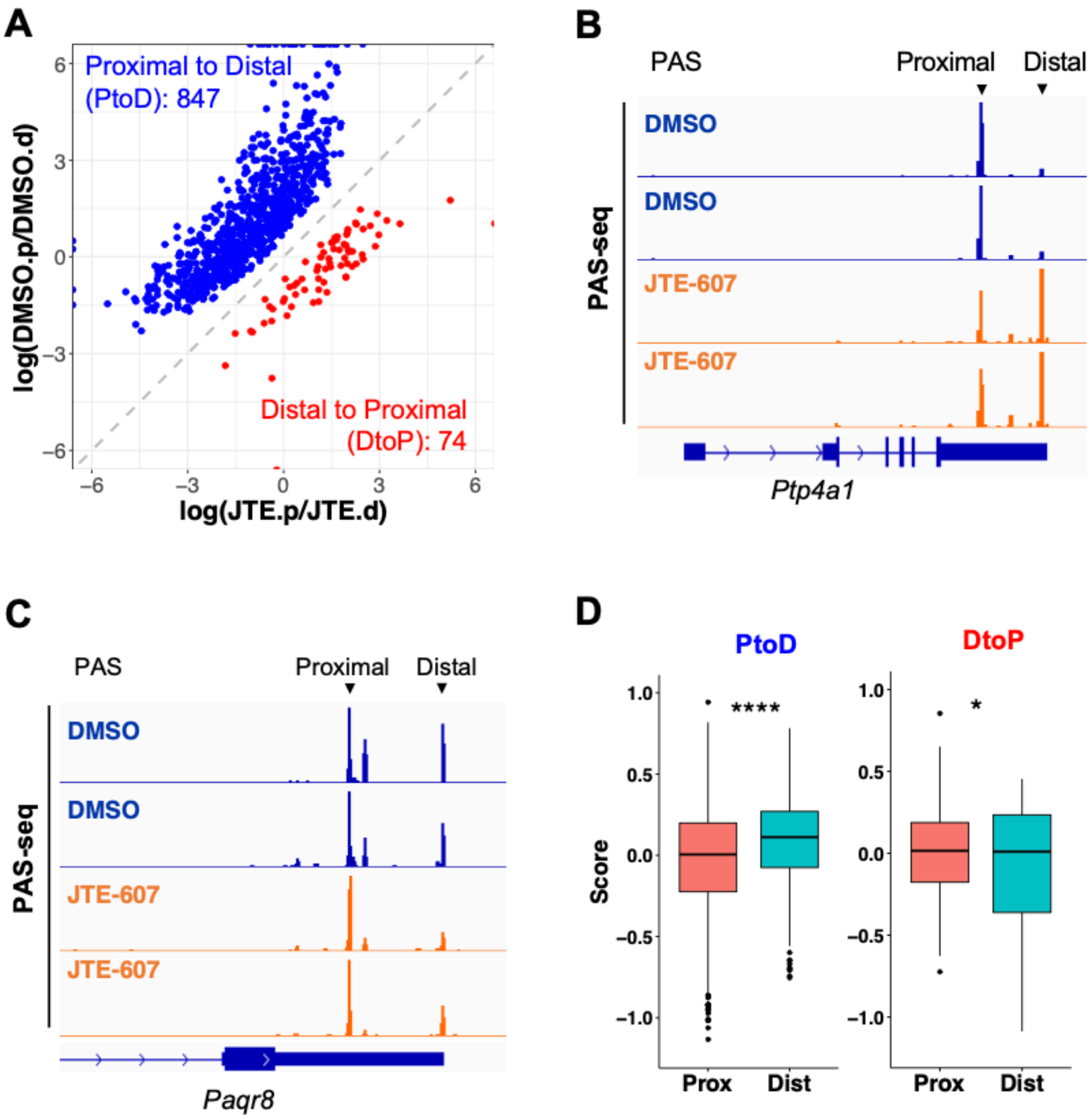

Figure 3.14. JTE-607-induced APA changes in cells are sequence-specific

(A) A scatter plot showing JTE-607-induced APA changes in cell.

(B-C) PAS-seq tracks of 2 example genes: *Ptp4a1* and *Paqr*8. Two replicates for each treatment are shown and the positions of the proximal and distal PASs are marked.

(D) Boxplots comparing the C3PO-predicted resistance scores for the proximal (Prox) and distal (Dist) PASs for the PtoD and DtoP genes. ****: p value < 0.0001; *: p value < 0.05 (t-test).

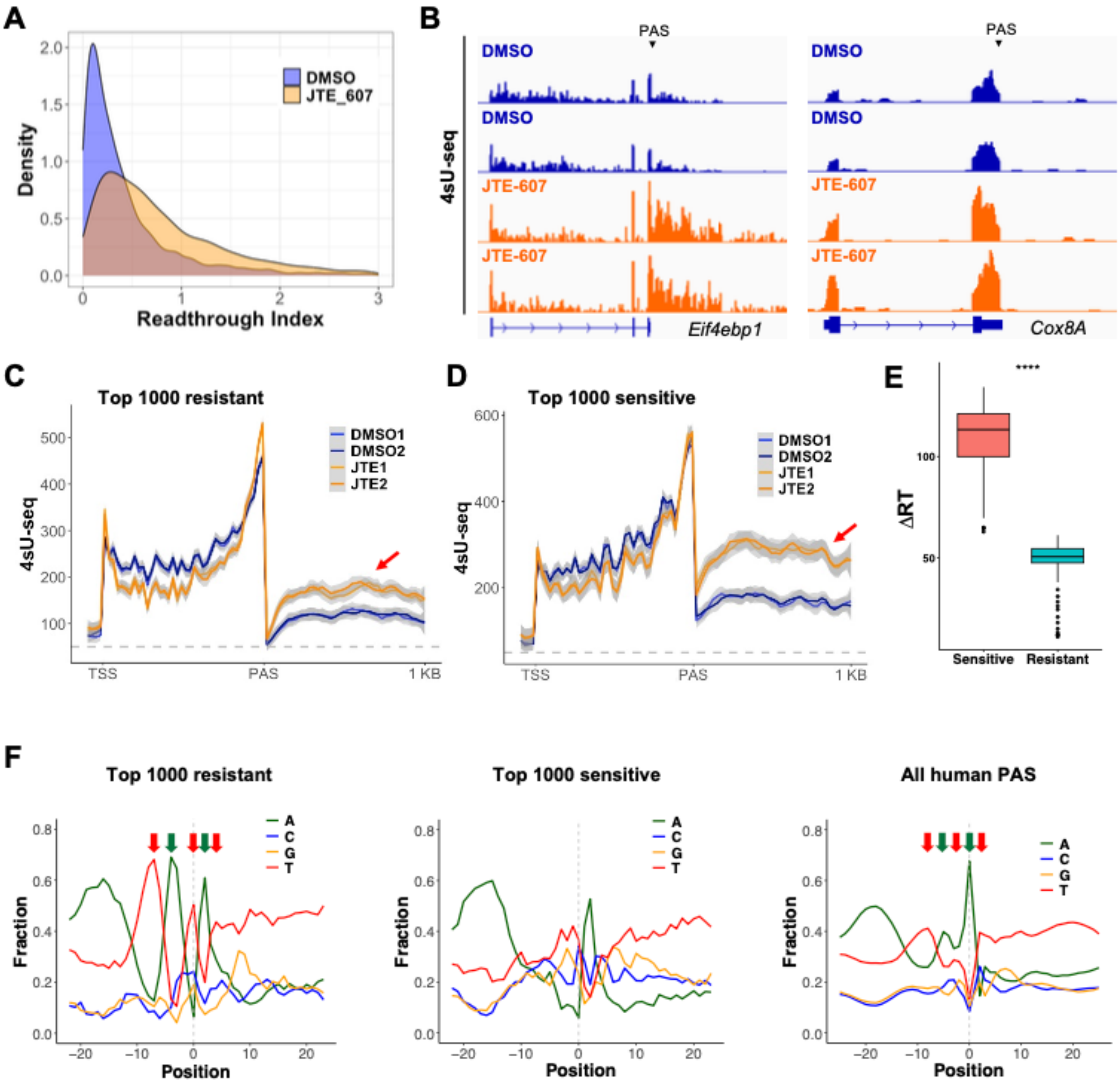
A
Density
2.0
1.5
1.0
0.5
0.0
DMSO
JTE_607
0
1
2
3
Readthrough Index
B
4sU-seq
PAS
DMSO
DMSO
JTE-607
JTE-607
Eif4ebp1
Cox8A
C
Top 1000 resistant
4sU-seq
500
400
300
200
100
DMSO1
DMSO2
JTE1
JTE2
TSS
PAS
1 KB
D
Top 1000 sensitive
4sU-seq
600
400
200
DMSO1
DMSO2
JTE1
JTE2
TSS
PAS
1 KB
E
****
ΔRT
100
50
Sensitive
Resistant
F
Top 1000 resistant
Top 1000 sensitive
All human PAS
Fraction
0.8
0.6
0.4
0.2
0.0
A
C
G
T
−20
−10
0
10
20
Position

Figure 3.15 JTE-607-mediated inhibition of mRNA 3’ processing in cells is sequence-specific

(A). A density plot of transcription readthrough index (read counts in the 1kb downstream region/read counts in gene body) for DMSO- and JTE-607-treated cells based on 4sU-seq data.

(B). 4sU-seq tracks for *Eif4ebp1* and *Cox8A* genes. Two replicates for DMSO and JTE-607 are shown. PAS positions are marked.

(C) Average normalized 4sU-seq signals for the genes with the top 1,000 most resistant PASs.

(D) Similar to C, but for the top 1,000 most sensitive PASs. Red arrow denotes region downstream of PAS.

(E) A blox plot compare the ΔRT (the difference in 4sU-seq signals in the 1kb region downstream of the PAS. ****: p value < 0.0001, Wilcoxan test.

(F) CS region nucleotide distribution for the top 1000 most resistant (left) and most sensitive (middle), and all human PASs (right). The T- and A-rich regions are marked by red and green arrows.

We then tested if C3PO can predict the transcription termination efficiency in JTE-607-treated cells. For comparison, we selected genes with the top 1000 resistant or sensitive PASs based on the C3PO predicted resistance scores. To avoid complications from neighboring genes, we selected genes that do not overlap with other genes in the 1kb downstream region for our analyses. The average normalized 4sU-seq signals at genes with the top 1000 resistant PASs showed that transcription terminated efficiently at these PASs in both DMSO- and JTE-607-treated cells and only modest change in RT levels was observed downstream of the PASs (Figure 3.15C, red arrow), suggesting that these PASs are indeed resistant to JTE-607. By sharp contrast, for genes with the top 1000 sensitive PASs, their global 4sU-seq signals revealed significantly higher RT in JTE-607-treated cells compared to DMSO-treated cells (Figure 3.15D, red arrow), suggesting that JTE-607 induced significant inhibition of mRNA 3′ processing at these PASs. The JTE-607-induced RT levels between the sensitive and resistant PASs were highly significant (Figure 3.15E, $p < 2.2 \times 10^{-16}$, Wilcoxon test). Together, our PAS-seq and 4sU-seq analyses suggest that JTE-607 inhibits mRNA 3′ processing and transcription termination in a sequence-dependent manner and that C3PO can predict the effect of JTE-607 on PAS selection and transcription termination.

Nucleotide composition of the resistant and sensitive human PASs revealed distinct patterns. JTE-607-resistant PASs have alternating U- and A-rich regions (Figure 3.15F, left panel) whereas the JTE-607-sensitive PASs are generally U/G-rich (Figure 3.15F, middle panel). These patterns are very consistent with the top resistant and sensitive PASs from our MPIVA screen (Figure 3.5F-G). Interestingly, the average nucleotide composition of the CS regions of all annotated human PASs also displayed alternating U- and A-rich regions (Figure 3.15F, right panel), suggesting that a significant portion of the human PASs are potentially resistant to

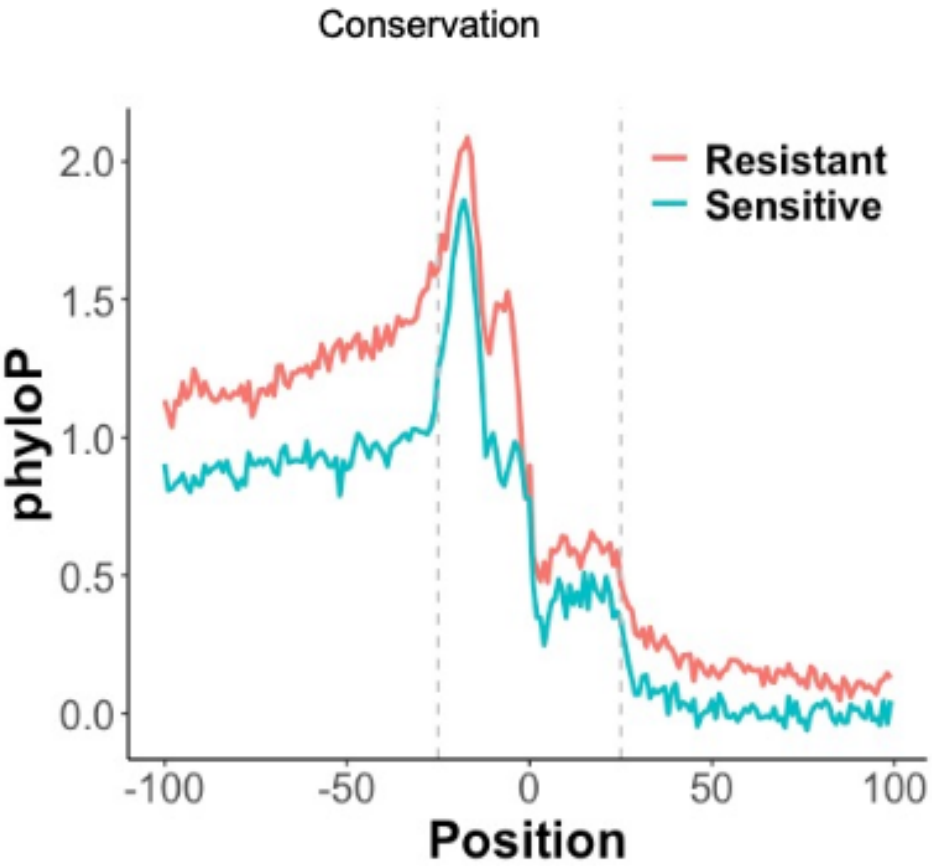

Figure 3.16 Conservation of JTE-607-sensitive and –resistant PASs

The phyloP sequence conservation score for both resistant and sensitive PASs across different species was calculated and plotted against nucleotide position of the CS. Position 0 is the YA (Y is U or C) cleavage position.

JTE-607. Finally, a comparison of the resistant and sensitive PASs revealed that the resistant PASs are more conserved than the sensitive PASs (Figure 3.16), indicating that the resistant PASs may be under greater selection pressure.

3.4 Discussion

In this study, we set out to characterize the mechanism of action for JTE-607, a novel inhibitor of the endonuclease for mRNA 3′ processing, CPSF73. Although CPSF73 is universally required for mRNA 3′ processing, we have unexpectedly discovered that Compound 2, the active form of JTE-607, inhibits the cleavage step of mRNA 3′ processing in a sequence-dependent manner both in vitro and in cells, and that the CS region sequence is a major determinant of Compound 2 sensitivity. We have comprehensively characterized the relationship between the CS region sequence and Compound 2 sensitivity using MPIVA coupled with machine learning. Our machine learning model C3PO can predict Compound 2 sensitivity based on CS sequence and the impact of JTE-607 on APA and transcription termination in human cells. Therefore, our study not only provided new insights into the mRNA 3′ processing machinery, but may also have important implications for the use of JTE-607 as a research and therapeutic tool. Furthermore, from a technological perspective, our approach described here should be broadly applicable to the studies of other small molecule modulators of gene expression.

What is the molecular mechanism for the sequence-specific sensitivity to Compound 2? Since both Compound 2 and the RNA in the CS region bind to CPSF73 at or near its active site (Sun et al. 2020; Ross et al. 2020), these interactions are most likely mutually exclusive (Figure

3.17). Thus, if a CS region RNA can bind to CPSF73 with a high affinity, it may out-compete Compound 2, rendering this PAS resistant to the drug (Figure 3.17, top panel). For low-affinity CS region RNA sequences, Compound 2 bound to CPSF73 near its active site can block access by the RNA due to its low affinity, thus inhibiting cleavage (Figure 3.17, bottom panel). Based on the structure of the histone mRNA cleavage complex (Sun et al. 2020), which contains CPSF73 as its endonuclease, CPSF73 binds to RNA substrates via a cleft between the β-CASP domain and the metallolactamase domain. However, this cleft can only accommodate a ~7 nt sequence, much shorter than the ~20 nt CS region that we identified. Thus, additional mRNA 3′ processing factors likely bind to the CS region as well. Potential candidates include CPSF100 and symplekin, which form the nuclease module, or mCF, with CPSF73 (Shi and Manley 2015; Sun et al. 2020). Indeed, the histone mRNA cleavage complex structure revealed that these proteins form an RNA-binding channel that can bind to ~20 nt sequence (Figure 3.17). Other mRNA 3′ processing factors can also be involved, including Fip1 and PAP. Fip1 is known to bind to U-rich sequences near the AAUAAA hexamer (Martin et al. 2012; Lackford et al. 2014). and the Compound 2-resistant CS sequences contain U-rich sequences (Figure 3.5F-G). Finally, an early biochemical study showed that PAP is required for in vitro cleavage of L3 PAS, but not for SVL and that the CS region sequences determine its PAP dependency (Ryner et al. 1989). Given the important roles for the CS region in determining both PAP dependency and Compound 2 sensitivity, it is possible that PAP is involved in binding to CS region sequences. Based on these results, we propose that CPSF73 and other mRNA 3′ processing factors form an RNA-binding channel that directly binds to the CS region RNA and that this channel has sequence specificity (Figure 3.17).

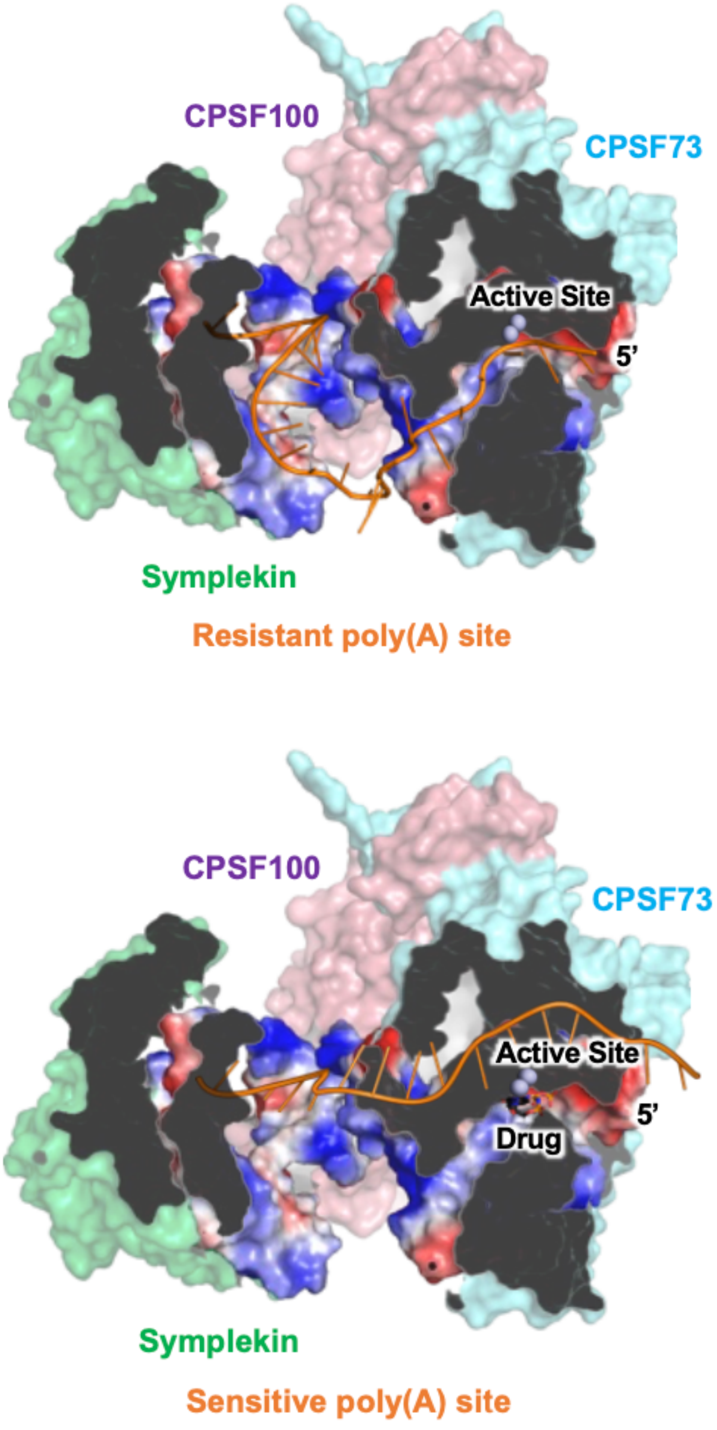

Figure 3.17 A model for sequence-specific inhibition of pre-mRNA 3’ processing by Compound 2

The nucleotide composition in the CS region has been conserved from yeast to human (Ozsolak et al. 2010; Liu et al. 2017; Derti et al. 2012). and this pattern is highly similar to that of the Compound 2-resistant PASs (Figure 3.15F). Additionally, our data suggests that Compound 2-resistant PASs are more evolutionarily conserved than the sensitive sites (Figure 3.16). It remains unclear what, if any, selection pressure can favor PASs that are resistant to a small molecule that is not present in most environments. We propose two possible models. First, Compound 2 activity may be similar to that of a chemical that is more universally found in cells. A number of small molecules, including inositol hexakisphosphate, can bind to and modulate the activities of molecular machinery in the gene expression pathway, such as the spliceosome (Wan et al. 2019), the Integrator complex (Lin et al. 2022), and the mRNA export factors (York et al. 1999). It is possible that a naturally occurring Compound 2-like small molecule can inhibit pre-mRNA 3′ processing and many PASs evolved to overcome such inhibition. Secondly, the CS region sequence may impact transcription termination independently of its effect on cleavage efficiency. According to our model, the resistant PASs interact with CPSF73 and other mRNA 3′ processing factors more strongly (Figure 3.17). Because the mRNA 3′ processing machinery is known to directly bind to RNA polymerase II (Richard and Manley 2009; Bentley 2005; Proudfoot 2016), such interaction could contribute to slowing down the polymerase, thus promoting termination. Thus, a subset of PASs may have evolved to stimulate transcription termination and the Compound 2 resistance is an unintended consequence of such evolution.

In addition to CS region sequence, we have provided evidence that the UPS sequence as well as RNA secondary structure may also contribute to Compound 2 sensitivity, albeit in a context-dependent manner (Figure 3.3A and Figure 3.13). UPS sequence can modulate CPSF73-CS region RNA interactions indirectly through associated protein factors. Alternatively, UPS could form secondary structures with the CS region, thus impacting its interactions with CPSF73 more directly. In fact, the mCF module has been shown to bind directly to double-stranded RNAs in the histone mRNA cleavage complex (Sun et al. 2020). Additionally, secondary structures are widespread in PAS regions and have been shown to modulate mRNA 3′ processing (Wu and Bartel 2017). Thus, it is possible that RNA secondary structures within a PAS could impact not only its cleavage efficiency, but also its Compound 2 sensitivity. Further studies are needed to fully elucidate the roles of UPS and RNA secondary structures in drug sensitivity.

Our results may have implications for understanding how JTE-607 specifically kills myeloid leukemia and Ewing's sarcoma cell lines. As mentioned earlier, JTE-607 is a pro-drug and is converted to Compound 2 by the cellular enzyme CES1 (Ross et al. 2020). Although cellular CES1 levels may contribute to the cell type specificity, previous studies showed that CES1 level is a poor predictor for JTE-607 sensitivity (Ross et al. 2020). Thus, the molecular basis for cell type-specific toxicity of JTE-607 remains unknown. Based on the results reported here, we propose two possible mechanisms for explaining the cell type-specific drug sensitivity. First, the potency for JTE-607-mediated inhibition of mRNA 3′ processing may be cell type-specific. Our model suggests that the drug sensitivity is determined by the interaction affinity between the CPSF73 and other mRNA 3′ processing factors and the CS region sequence. If cell type-specific mechanisms can modulate the specificity of this interaction, they can alter JTE-607 sensitivity globally. This could result from cell type-specific expression levels or post-translational modification of CPSF73 and other mRNA 3′ processing factors that bind to the CS region. For example, Fip1 levels are known to change during differentiation (Lackford et al. 2014; Schwich et al. 2021, 3). Symplekin, CPSF100, and PAP are known to be sumoylated and/or phosphorylated (Vethantham et al. 2008, 2007; Colgan et al. 1996). It will be important to determine if these factors display different expression levels or post-translational modifications between JTE-607-sensitive and -resistant cell types. Alternatively, the sequence specificity of JTE-607 is similar among different cell types. However, myeloid leukemia and Ewing's sarcoma cells may be uniquely dependent on one gene or a subset of genes whose PASs are highly sensitive to JTE-607. For example, a recent study identified PDXK, an enzyme in the vitamin B6 metabolism pathway, as a unique acute myeloid leukemia dependency gene (Chen et al. 2020). If the PASs of such dependency genes are sensitive to JTE-607, the expression of these genes would be repressed by JTE-607 treatment, leading to cell death in specific cell types. Further studies are needed to distinguish between these models and the results will have significant implications on how to improve the efficacy of this compound as a potential anti-cancer and anti-inflammation therapy.

Table 3.1 Activity and IC_{50} of all PASs plotted in Figure 3.1D

PAS Name	Activity	IC_{50} (μM)
L3	28.04	0.8
SVL	36.12	100.2
L3-SVL-Up	51.99	2.2
L3-SVL-Down	54.62	6.7
L3-SVL-CS	44.14	47.8
SVL-L3-Up	31	39.5
SVL-L3-Down	32.46	89.6
SVL-L3-CS	14.2	0.8
SVL-L3-CS2	13.36	0.6
SVL-L3-CS3	17.67	0.5
SVL-L3-CS4	9.15	5.5
L3m3	4.76	0.7
SVL230	10.75	45.2
L3-SVL230-CS	20.94	0.9
SVL230-L3-CS	0.15	20.2
PerPAS	15.04	1.0
PerPAS-dA	17.57	0.4
PerPAS-GUGUm	0.53	2.3
PerPAS-UGUAm	3.37	0.7
BASP1	6.08	0.5
SAU5	10.59	0.9
HO-1	1.38	0.8
ACTB	19.63	13.9
UCK2	5.7	0.5
CBX6	2.65	0.3
GAPDH	38.09	0.5
ICP27	8.03	0.5
bGH	26.08	38.1
L3-Sen1	15.5	0.4
L3-Sen52	5.49	0.2
L3-Sen84	17.65	0.2
L3-Rst14	46.65	42.9
L3-Rst34	33.81	8.3
L3-Rst52	41.99	16.4
SVL-Sen3	9.03	0.2
SVL-Sen38	18.72	0.6
SVL-Sen127	10.8	0.4
SVL-Rst303	18.46	72.4
L3 UA to CA	27.8	2.2
SVL CA to UA	20.1	148.0

Table 3.2 R^2 for various machine learning architectures tested

Model	Test R^2: 0.5 µM	Test R^2: 2.5 µM	Test R^2: 12.5 µM
C3PO	0.31	0.552	0.7
CNN - 4 epochs	0.282	0.53	0.682
CNN - 4 epochs, 12.5 µM only	N/A	N/A	0.683
CNN - 5 epochs	0.301	0.542	0.687
CNN - 5 epochs, 12.5 µM only	N/A	N/A	0.689
CNN - 6 epochs	0.291	0.523	0.687
CNN - 6 epochs, 12.5 µM only	N/A	N/A	0.704
CNN - 7 epochs	0.3	0.543	0.691
CNN - 7 epochs, 12.5 µM only	N/A	N/A	0.687
CNN - 8 epochs	0.306	0.536	0.684
CNN - 8 epochs, 12.5 µM only	N/A	N/A	0.688
CNN - validation early stop	0.314	0.546	0.694
CNN - validation early stop, 12.5 µM only	N/A	N/A	0.691
CNN - 4 epochs, Trial 2	0.311	0.54	0.688
CNN - 4 epochs, 12.5 µM only, Trial 2	N/A	N/A	0.682
CNN - 5 epochs, Trial 2	0.305	0.539	0.687
CNN - 5 epochs, 12.5 µM only, Trial 2	N/A	N/A	0.695
CNN - 6 epochs, Trial 2	0.307	0.548	0.699
CNN - 6 epochs, 12.5 µM only, Trial 2	N/A	N/A	0.701
CNN - 7 epochs, Trial 2	0.302	0.54	0.686
CNN - 7 epochs, 12.5 µM only, Trial 2	N/A	N/A	0.69
CNN - 8 epochs, Trial 2	0.296	0.538	0.69
CNN - 8 epochs, 12.5 µM only, Trial 2	N/A	N/A	0.69
CNN - validation early stop, Trial 2	0.307	0.539	0.688
CNN - validation early stop, 12.5 µM only, Trial 2	N/A	N/A	0.696
CNN - 6 epochs, Trial 1 out of 10	0.292	0.544	0.687
CNN - 6 epochs, Trial 1 out of 10, 12.5 µM only	N/A	N/A	0.701
CNN - 6 epochs, Trial 2 out of 10	0.299	0.538	0.696
CNN - 6 epochs, Trial 2 out of 10, 12.5 µM only	N/A	N/A	0.69
CNN - 6 epochs, Trial 3 out of 10	0.307	0.553	0.699
CNN - 6 epochs, Trial 3 out of 10, 12.5 µM only	N/A	N/A	0.701
CNN - 6 epochs, Trial 4 out of 10	0.317	0.541	0.689
CNN - 6 epochs, Trial 4 out of 10, 12.5 µM only	N/A	N/A	0.694
CNN - 6 epochs, Trial 5 out of 10	0.31	0.548	0.695
CNN - 6 epochs, Trial 5 out of 10, 12.5 µM only	N/A	N/A	0.69
CNN - 6 epochs, Trial 6 out of 10, 12.5 µM only	N/A	N/A	0.69
CNN - 6 epochs, Trial 7 out of 10	0.277	0.514	0.672
CNN - 6 epochs, Trial 7 out of 10, 12.5 µM only	N/A	N/A	0.695
CNN - 6 epochs, Trial 8 out of 10	0.304	0.546	0.693
CNN - 6 epochs, Trial 8 out of 10, 12.5 µM only	N/A	N/A	0.692

CNN - 6 epochs, Trial 9 out of 10	0.295	0.545	0.697
CNN - 6 epochs, Trial 9 out of 10, 12.5 µM only	N/A	N/A	0.695
CNN - 6 epochs, Trial 10 out of 10	0.305	0.55	0.697
CNN - 6 epochs, Trial 10 out of 10, 12.5 µM only	N/A	N/A	0.687
CNN - hyperband training	0.198	0.441	0.585
CNN - hyperband training, 12.5 µM only	N/A	N/A	0.586
ResNet	0.232	0.541	0.681
ResNet - 7 epochs	0.216	0.497	0.632
ResNet - 12221 dilations	0.221	0.494	0.647
ResNet - 11111 dilations	0.221	0.525	0.657
ResNet - cleavage length 27	0.229	0.55	0.686
ResNet - cleavage length 27, 75% loss for Compound 2 sensitivity predictions	0.256	0.543	0.68

Table 3.3 Log (12.5 µM/DMSO) and C3PO predicted score for all PASs plotted in Figure 3.10C

PAS Name	Log (12.5 µM/DMSO)	C3PO Predicted Score
L3	-0.67	-0.72
SVL	-0.14	0.34
BASP1	-0.71	0.15
SAU5	-0.54	0.21
PerPAS	-0.99	-0.61
PerPAS-dA	-2.13	-0.88
PerPAS-GUGUm	-0.83	-0.60
HO-1	-0.73	0.15
L3m3	-1.00	-0.81
ACTB	-0.28	0.31
UCK2	-0.60	0.18
CBX6	-0.63	-0.10
GAPDH	-0.89	0.31
ICP27	-0.44	-0.42
bGH	-0.22	0.02
SVL-L3-CS2	-0.52	0.04
SVL-L3-CS3	-0.29	-0.11
SVL-L3-CS4	-0.18	0.74
L3-Sen1	-1.62	-1.02
L3-Sen52	-1.47	-1.49
L3-Sen84	-1.92	-1.92
L3-Rst14	-0.23	0.94
L3-Rst34	-0.26	0.96
L3-Rst52	-0.29	0.96
SVL-Sen3	-1.18	-1.15
SVL-Sen38	-1.24	-0.94
SVL-Sen127	-1.16	-0.50
SVL-Rst27	-0.18	1.04
SVL-Rst55	-0.17	0.30
SVL-Rst303	-0.21	0.55

Table 3.4 Sequence of all in vitro tested PAS in this study

PAS NAME	DNA SEQUENCE
L3	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGGCAAATGTTTTTATTTGTACACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
SVL	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
L3-SVL-Up	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAGGCAAATGTTTTTATTTGTACACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGA
L3-SVL-Down	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGGCAAATGTTTTTATTTGTACATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
L3-SVL-CS	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAACAAGTTAACAACAACAATTGCACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGA
SVL-L3-Up	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-L3-Down	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGA
SVL-L3-CS	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAGGCAAATGTTTTTATTTGTACATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-L3-CS2	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAATTTGTACATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-L3-CS3	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTTGTTTTTAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-L3-CS4	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACATTTATTTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA

L3m3	TTCTTTTTGTCACTTGAAAAACATGGGAAAATAATGGGCTAGGAGACACTTTCAATAAAGGCAAATGTTTTATTTTTACACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
SVL230	AGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCCAGATCCCCGGGTACCGAGCTC
L3-SVL230-CS	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGGCAAAAACAACAACAATTGCACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
SVL230-L3-CS	AGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTTGTTTTTATTTGTACATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCCAGATCCCCGGGTACCGAGCTC
PerPAS	TTTTTTTTTTTGTAAATTAATTTTTAATAAAGTTGTTTTTTACACGTTGTCTTGTGTGTCGTCTTTTTTTTCAGC
PerPAS-dA	TTTTTTTTTTTGTAAATTAATTTTTAATAAAGTTGTTTTTTGCGCGTTGTCTTGTGTGTCGTCTTTTTTTTCAGC
PerPAS-GUGUm	TTTTTTTTTTTGTAAATTAATTTTTAATAAAGTTGTTTTTTACACGTTGTCTTCTCTCTCGTCTTTTTTTTCAGC
PerPAS-UGUAm	TTTTTTTTTTTCTACATTAATTTTTAATAAAGTTGTTTTTTACACGTTGTCTTGTGTGTCGTCTTTTTTTTCAGC
BASP1	TGAAAGGGAAAGCCTAGCAAGTCTATTAATAAGCTCACTTCCCATTTATCCCAGTGTACCTGGAGCATTAAGCTAAGACGTTCATCCACAGGCTTAAAAACTTACATCAAGCACTACTGAACTTTACAAGCTGGAATAAACAATGCCTACTAAATAAAAGATTTATAAAATTGTTCTGTCTTATTTTTGTGATCTCTTGTAAATGTTTTTTTTTTTTTTTTTTTTTAAATATCCAAAGAAGACCTGTGAACTATTATTTGTCAGAAGCAATTGCCCTTGGTATCTGATTCTGTTGAAAGAA
SAU5	GCCGCCCAACCCGAGCGACCTTCCCCTCCCACTTCCCCCCCCCTACACACCAACTCCGCCCTCGCCGTCTTGGCCGTGCGCGGCCCCGTGCGTCCGTCTCAATAAAGCCAGGTTAAATCCGTGACGTGGTGTGTTTGGCGTGTGTCTCTGAAATGGCGGAAACCGACATGCAAATGGGATTCATGGACATGTTACACCCCCCTGAC
HO-1	GGCACTGTGGCCTTGGTCTAACTTTTGTGTGAAATAATAAACAACATTGTCTGATAGTAGCTTGAAGTAGTTTTCATGGGCTTTGTTATTCTTGGGGAACTGACCTTTTCCTCCCTGGTTTCTTGCGTGCTCGGTAGGA

ACTB	TTTTTTGTCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCCCTGGGAGTGGGTGGAGGCAGCCAGGGCTTACCTGTACACTGACTTGAGACCAGTTGAATAAAAGTGCACACCTTAAAAATGAGGCCAAGTGTGACTTTGTGGTGTGGCTGGGTTGGGGGCAGCAGAGGGTGAACCCTGCAGGAGGGTGAACCCTGCAAAAGGGTGGGGCAGTGGGGGCCAAC
UCK2	CCCCCCTTTTAAGATGCTTGCTCCTCTCCCTTTTCTTTTTACCACCCTACCTTTATTGTTAGTGGTTACAAAGTGACCACATATTATGTACTTTGCTGTAAATAAAGACAGACAAAAAGGCTCTCGCCTTCTGTGTGATGCTTGGCCCAGAGCAGCGACCGAATCCTGGCTGTGTGGCCCAAGTGGCTCAGGAAGGGCCATGCTGTGCATGTGTGGTGTAGA
CBX6	GGGTCTGTGCCGATTACTCTGTCTTGTACGTTTGTTCTGCTGCTCTTCAATATTGTATCAACGCCAGGAAAGGGGGGTGAAAAGCCTCTTTTACCCCCCAAATAAATTGTCACATTCCGAAGCTGAGGCCTAGCCCCTAGGTTGGGGTGTGTCTGTGTCTTCTTCCAGCTGTGACTGGCTTTTCAAAAGTAGCAGGCCCATGTCCCTCCAGTGACAGGTGAAGAGGGG
GAPDH	TCAGTCCCCCACCACACTGAATCTCCCCTCCTCACAGTTGCCATGTAGACCCCTTGAAGAGGGGAGGGGCCTAGGGAGCCGCACCTTGTCATGTACCATCAATAAAGTACCCTGTGCTCAACCAGTTACTTGTCCTGTCTTATTCTAGGGTCTGGGGCAGAGGGGAGGGAAGCTGGGCTTGTGTCAAGGTGAGACATTCTTGCTGGGGAGGGACCTGGTATGT
ICP27	GTGTTCGAGTCGTGTCTGCGAGTTGACGGCCAGTCACATCGTCGCCCCCCCGTACGTGCACGGCAAATATTTTTATTGCAACTCCCTGTTTTAGGTACAATAAAAACAAAACATTTCAAACAAATCGCCCCTCGTGTTGTCCTTCTTTGCTCATGGCCGGCGGGGCGTGGGTCACGGCAGATGGCGGGGGTGGGCCCGGCG
bGH	CTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGG
L3-Sen1	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAATGTGATTGTTTCAATCGGAGATTGTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
L3-Sen52	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAACAATGTGCTGTTCAAAGGCGGTGGCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
L3-Sen84	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGCGAAATGTTGTTAATGTGCCCGCG

	TCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
L3-Rst14	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAAAGGTTAACGCTCATATGGTTCGTTTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
L3-Rst34	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAAACCGTTAACGCTATAGTTGGCTGGTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
L3-Rst52	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGACGTTGAACTTCATAATCGTGCCATCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
SVL-Sen3	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAATGCGTAGGCAGGTGTCGTATCGATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-Sen38	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAAACATGTCGTGCATTTGTTTCATTGATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-Sen127	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAAATCTAATGTGTAAAAGGTTTAAGTATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-Rst27	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACGAGTTAACGCTACTTTCGGTTTCTATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-Rst55	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAGTGCGGTAACGCAGAATTTTGTAATATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
SVL-Rst303	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAAGGGGTTAACTACATGACAAGGTGACATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
L3 UA to CA	TTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAAGGCAAATGTTTTCATTTGTACACTCTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGAGGTACCGAGCTC
SVL CA to UA	ATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAATAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA

Table 3.5 Oligonucleotides and synthesized DNA used in this study

OLIGO NAME	DNA SEQUENCE
For PAS cloning	
L3-F	ACAGGATCCTTCTTTTTGTCACTTGAAAAACATGTA
L3-R	ACAGAATTCTCGCAGACGGCAAG
SVL230-F	ACAGGATCCAGACATGATAAGATACATTGATGAG
SVL230-R	ACAGAATTCGAGCTCGGTACCCG
SVL-F	ACAGGATCCATGCTTTATTTGTGAAATTTGTGATGCTATTG
SVL-R	ACAGAATTCTACTTGCTTTAAAAAACCTCCCACAC
L3-SVL Up-F	GTAACCATTATAAGCTGCAATAAAGGCAAATGTTTTTA
L3-SVL Up-R	TAAAAACATTTGCCTTTATTGCAGCTTATAATGGTTAC
L3-SVL Down-F	CAAATGTTTTTATTTGTACATTCATTTTATGTTTCA
L3-SVL Down-R	TGAAACATAAAATGAATGTACAAATAAAAACATTTG
SVL-L3 Up-F	GTACTAGGAGACACTTTCAATAAACAAGTTAACAAC
SVL-L Up-R	GTTGTTAACTTGTTTATTGAAAGTGTCTCCTAGTAC
SVL-L3 Down-F	GTTAACAACAACAATTGCACTCTCGGGTGATTATTTA
SVL-L3 Down-R	TAAATAATCACCCGAGAGTGCAATTGTTGTTGTTAAC
L3-SVL CS-F	CAAGTTAACAACAACAATTGCACTCTCGGGTGATTATTTAC
L3-SVL CS-R	CAATTGTTGTTGTTAACTTGTTTATTGAAAGTGTCTCC
SVL-L3 CS-F	GGCAAATGTTTTTATTTGTACATTCATTTTATGTTTCAG
SVL-L3 CS-R	TACAAATAAAAACATTTGCCTTTATTGCAGCTTATAATG
SVL-L3 CS2-F	ATTTGTACATTCATTTTATGTTTCAGGTTCAG
SVL-L3 CS2-R	TGTTGTTAACTTGTTTATTGCAGC
SVL-L3 CS3-F	TGTTTTTAACAATTGCATTCATTTTATGTTTCAG
SVL-L3 CS3-R	AACTTGTTTATTGCAGCTTATAATG
SVL-L3 CS4-F	TTTATTTGCATTCATTTTATGTTTCAGGTTCAG
SVL-L3 CS4-R	TGTTAACTTGTTTATTGCAGC
L3m3-F	CAAATGTTTTTATTTTTACACTCTCGGGTG
L3m3-R	CACCCGAGAGTGTAAAAATAAAAACATTTG
PerPAS-F	ACATCTAGATTTTTTTTTTTGTAAATTAATTTTTAATAAAGTTGTTTTTT
PerPAS-R	ACACTCGAGGCTGAAAAAAAAGACGACACACAAGACAACGTGTAAAAAACAACTTT
PerPAS-mut-F	ACATCTAGATTTTTTTTTTTGTAAATTAATTTTTAACAAAGTTGTTTTTT
PerPAS-dA-R	ACACTCGAGGCTGAAAAAAAAGACGACACACAAGACAACGCGCAAAAAACAACTTT
PerPAS-UGUAmut-R	ACATCTAGATTTTTTTTTTTCTACATTAATTTTTAATAAAGTTGTTTTTT
PerPAS-GUGUmut-R	ACACTCGAGGCTGAAAAAAAAGACGAGAGAGAAGACAACGTGTAAAAAACAACTTT

BASP1-F	AGTCTAGATGAAAGGGAAAGCCTAGC
BASP1-R	TTCTCGAGTTCTTTCAACAGAATCAG
SAU5-F	ACATCTAGAGCCGCCCAACCCGAGCGACCT
SAU5-R	ACACTCGAGGTCAGGGGGGTGTAACATGTCCA
HO1-F	ACTCTAGAGGCACTGTGGCCTTGGTCTAA
HO1-R	ATCTCGAGtCCTACCGAGCACGCAAGAA
ACTB-F	ACATCTAGATTTTTTGTCCCCCAACTTGAG
ACTB-R	ACACTCGAGGTTGGCCCCCACTGCCCCAC
UCK2-F	ACATCTAGACCCCCCTTTTAAGATGCTTG
UCK2-R	ACACTCGAGTCTACACCACACATGCACAG
CBX6-F	ACATCTAGAGGGTCTGTGCCGATTACTCT
CBX6-R	ACACTCGAGCCCCTCTTCACCTGTCACTG
GAPDH-F	ACATCTAGATCAGTCCCCCACCACACTGA
GAPDH-R	ACACTCGAGACATACCAGGTCCCTCCCCA
ICP27-F	ACATCTAGAGTGTTCGAGTCGTGTCTGCGAG
ICP27-R	ACACTCGAGCGCCGGGCCCACCCCCGCCATCTGCCGTGACCCAC
bGH-F	ACAGGATCCCTGTGCCTTCTAGTTGCCAG
bGH-R	ACAGAATTCCCATAGAGCCCACCGCATC
L3 Sen1-F	ATGTGATTGTTTCAATCGGAGATTGTCGGGTGATTATTACCCCCCAC
L3 Sen52-F	CAATGTGCTGTTCAAAGGCGGTGGCTCGGGTGATTATTTACCCCCCAC
L3 Sen84-F	GCGAAATGTTGTTAATGTGCCCGCGTCGGGTGATTATTTACCCCCCAC
L3 Rst14-F	AAGGTTAACGCTCATATGGTTCGTTTCGGGTGATTATTACCCCCCAC
L3 Rst34-F	AACCGTTAACGCTATAGTTGGCTGGTCGGGTGATTATTTACCCCCCAC
L3 Rst52-F	GACGTTGAACTTCATAATCGTGCCATCGGGTGATTATTTACCCCCCAC
L3 linear-R	TTTATTGAAAGTGTCTCCTAGTACATTATTTTTAC
SVL Sen3-F	CAATGCGTAGGCAGGTGTCGTATCGATTTTATGTTTCAGGTTCAGGGGGAG
SVL Sen38-F	AACATGTCGTGCATTTGTTTCATTGATTTTATGTTTCAGGTTCAGGGGGAG
SVL Sen127-F	AATCTAATGTGTAAAAGGTTTAAGTATTTTATGTTTCAGGTTCAGGGGGAG
SVL Rst27-F	CGAGTTAACGCTACTTTCGGTTTCTATTTTATGTTTCAGGTTCAGGGGGAG
SVL Rst55-F	GTGCGGTAACGCAGAATTTTGTAATATTTTATGTTTCAGGTTCAGGGGGAG
SVL Rst303-F	GGGGTTAACTACATGACAAGGTGACATTTTATGTTTCAGGTTCAGGGGGAG
SVLst linear-R	TTTATTGCAGCTTATAATGGTTACAAATAAAGC

L3 CAmt-F	CATTTGTACACTCTCGGGTGATTATTTAC
L3 CAmt-R	AAAACATTTGCCTTTATTGAAAGTGTC
SVLst UAmt-F	TAACAATTGCATTCATTTTATGTTTCAGGTTC
SVLst UAmt-R	TTGTTAACTTGTTTATTGCAGCTTATAATG
For MPIVA N23 screen	
T7-L3 N23	TAATACGACTCACTATAGGGATAATTTCTTTTTGTCACTTGAAAAACATGTAAAAATAATGTACTAGGAGACACTTTCAATAAANNNNNNNNNNNNYANNNNNNNNNNNTCGGGTGATTATTTACCCCCCACCCTTGCCGTCTGCGA
T7-SVL N23	TAATACGACTCACTATAGGGATAATATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAANNNNNNNNNNNYANNNNNNNNNNNNATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTTTTTAAAGCAAGTA
L3 N23 Gibson-F	CTATAGGGCGAATTGGAGCTCTTCTTTTTGTCACTTGAAAAACATGTA
L3 N23 Gibson-R	GTATCGATAAGCTTGATATCGAATTCTCGCAGACGGCAAGG
SVL N23 Gibson-F	CTATAGGGCGAATTGGAGCTCATGCTTTATTTGTGAAATTTGTGATGC
SVL N23 Gibson-R	GTATCGATAAGCTTGATATCGAATTCTACTTGCTTTAAAAAACCTCCCACAC

3.5 Methods

Cell Culture and JTE-607 Treatment Condition

HepG2 cell line were cultured in Dulbecco’s modified Eagle medium (DMEM) supplemented with 10% (v/v) fetal bovine serum (FBS). Cells were incubated at 37℃ in a 5% (v/v) CO_2-enriched incubator. Suspension HeLa S3 cell (a kind gift from Dr. Bert Semler, UC Irvine) was maintained in Joklik Modified MEM (JEME) supplemented with 2.4 mM sodium bicarbonate and 8% (v/v) newborn calf serum (NCS) in a spinner flask at 37˚C with ambient CO_2. For JTE-607 treatment, 20 µM final concentration of JTE-607 (Tocris) in neat DMSO was added to the cell culture media and incubated at 37˚C for 4 hours.

Large-scale HeLa nuclear extract (NE)

Large-scale HeLa nuclear extract (NE) was made as previously described (Abmayr et al. 2006) with minor modifications. Briefly, 10 liters of spinner HeLa cells were pelleted by centrifugation. The cells were swelled on ice using hypotonic buffer A (10 mM HEPES-NaOH pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 10 mM 2-Mercaptoethanol) and then dounce homogenized with 15 strokes using a type B pestle. Each stroke involves a 30-second motion containing one 15-second up and one 15-second down motion. Cell lysis was closely monitored by mixing a small aliquot of cells with trypan blue and observing under light microscope. Dounce homogenization was stopped when ~85% cell lysis was achieved. Nuclei were pelleted, extracted with high salt buffer C (20 mM HEPES-NaOH pH 7.9, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 25% glycerol, 10 mM 2-Mercaptoethanol, 0.5 mM PMSF) freshly supplemented with 1X

Halt proteinase inhibitor cocktail (Thermo) at 4˚C for one hour with constant rotation. The extracted nuclei were pelleted, and the supernatant (NE) was dialyzed twice against 60 volume of buffer D100 (20 mM HEPES-NaOH pH 7.9, 100 mM KCl, 1 mM $MgCl_2$, 0.2 mM EDTA, 10% glycerol 10 mM 2-Mercaptoethanol, 0.5 mM PMSF) at 4˚C for 1.5 hours each time. After dialysis, the NE was aliquoted, flash frozen on dry ice and stored at -80˚C until use.

In vitro Cleavage Assay

All PASs were cloned into the pBlueScript II KS+ vector. RNA substrates were synthesized by run off in vitro transcription (IVT) using T7 polymerase (NEB) in the presence of [α-^{32}P]-UTP according to the manufacture's protocol. For in vitro cleavage reaction with Compound 2, the NE was pre-incubated with 10% DMSO or various concentration of Compound 2 (0.1, 0.5, 2.5, 12.5, 62.5, 100 μM) in 10% DMSO for 30 minutes on ice before the other components were added. Each in vitro cleavage reaction is a 10μl reaction containing 20 cps radiolabeled pre-mRNA, 44% (v/v) HeLa NE, 8.8 mM HEPES-OH (pH 7.9), 44 mM KCl, 0.44 mM $MgCl_2$, 0.2 mM 3′-dATP (Sigma), 2.5% (v/v) polyvinyl alcohol (PVA), 40 mM creatine phosphate, 4 mM 2-Mercaptoethanol, and 1% (v/v) DMSO or Compound 2. Cleavage was carried out for 90 minutes at 30˚C. Proteinase K digestion mix (30 mM Tris-HCl pH 7.9, 10 mM EDTA, 1% SDS, 0.1 μg/μl proteinase K, 0.05 μg/μl yeast tRNA) was then added to halt the reaction and the samples were incubated at 37˚C for 15 min. RNA was then phenol chloroform extracted and resolved on an 8% Urea-PAGE at 800 V for 45 minuets in TBE. Gel was then transferred to a filter paper, dried at 80˚C for 30 minutes, exposed to a phosphoscreen overnight and visualized by phosphorimaging. IC_{50} was calculated using the equation: [Inhibitor] vs. normalized response -- Variable slope on Prism.

We have found that this assay is very sensitive to the strength of RNA radioactivity and freshness of NE. It is recommended that freshly purchased [α-^{32}P]-UTP (less than a week old) and NE made and stored at -80˚C for fewer than two months to be used for this assay.

Electrophoretic mobility shift assay (EMSA)

NE was pre-incubated with DMSO or Compound 2 as described in the in vitro cleavage assay. Gel shift is performed in a 10µl reaction containing 20 cps radiolabeled RNA, 1 mM ATP, 20 mM creatine phosphate, 10 µg/µl yeast tRNA, 44% HeLa NE, and 1% (v/v) DMSO or Compound 2. The reaction mixture was incubated for 20 minutes at 30˚C and immediately cooled on ice for 2 minutes. Heparin was added to 0.4 µg/µl and the reaction was incubated for an additional 5 minutes on ice. 5µl of the reaction was resolved on 4% native PAGE in 1x Tris-Glycine running buffer (pH 8.3) at 100V for 4 hours in an ice bath. Gel was dried and visualized the same as in vitro cleavage assay described above.

Massively Parallel in vitro Assay (MPIVA)

<u>Cloning</u>

L3 and SVL containing 23 random nucleotides CS spanning YA cleavage position was purchased from IDT as ssDNA oligo and PCR amplified to generate dsDNA. The dsDNA library was cloned into pBlueScript II KS+ vector by Gibson Assembly (NEB) and electroporated into ElectroMAX DH5α (Thermo). Plasmid library size, structure, and diversity were determined as previously described(Bogard et al. 2019).

Coupled in vitro Cleavage and Polyadenylation Assay

RNA libraries were synthesized by run off IVT using T7 polymerase (NEB) according to the manufacture's protocol followed by treatment with RQ1 DNase (Promega) to remove DNA template. The RNA pool was purified by phenol chloroform extraction and was either polyadenylated (for input) or 3′-dATP blocked (for DMSO and Compound 2 treated) by *E. coli* PAP (NEB). The RNAs were then undergo a coupled cleavage and polyadenylation assay in multiple 600μl reactions containing 6 pmol RNA, 44% (v/v) HeLa NE, 8.8 mM HEPES-OH (pH 7.9), 44 mM KCl, 1.44 mM $MgCl_2$, 1 mM ATP, 2.5% (v/v) polyvinyl alcohol (PVA), 20 mM creatine phosphate, 4 mM 2-Mercaptoethanol, and either 1% DMSO or 0.5 μM, 2.5 μM, 12.5 μM Compound 2 in DMSO. The reaction mixture was incubated for 90 min at 30˚C, proteinase K digested as described above for regular in vitro cleavage assay, except that proteinase K was raised to 3μg/μl, and then phenol chloroform extracted.

MPIVA Sequencing Library Construction

The phenol chloroform extracted RNA from previous step weas further purified to select for polyadenylated RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) and reverse transcribed using SuperScript III reverse transcriptase (Invitrogen) with an anchored oligo dT primer. Library cDNA was beads purified (Beckman Coulter) and amplified using a library-specific forward primer and reverse primer containing Illumina adaptor sequences and a region that matches part of the sequence added during RT. The amplified libraries were resolved on a 2.5% low melting point agarose gel and extracted.

MPIVA RNA-seq read alignment

All MPIVA read 1 and read 2 FASTQ files were merged using bbmerge V38 using option ` maxloose=t` (Bushnell et al. 2017). Untreated RNA-seq reads were used to establish the sequences of the full randomized PAS region contained in the IVT pool. All sequences of the 25 nt randomized region were clustered using starcode version 1.4 (Zorita et al. 2015) to account for sequencing errors and determine consensus sequences of this randomized region. The next steps are to enable assignment of expected cleaved RNAs to a unique 25 nt randomized region. The expected cleaved lengths of the 25 nt consensus sequences for L3 and SVL backbones (13 nt and 12 nt, respectively) were used as unique identifiers of the full randomized region. If any of these identifiers were not unique within L3 and SVL libraries, respectively, then these sequences were not used in subsequent analyses.

Next, RNA-seq reads from DMSO and drug-treated libraries were locally aligned against the shared 5′ region of the reporter constructs to determine the beginning of the randomized region. The part of the RNA-seq read containing the randomized region and shared 3′ region was locally aligned against the list of consensus sequences. Only reads with a unique alignment to a single consensus sequence were kept, and cut sites were also determined from this alignment. Additional checks were performed to ensure cut sites are not misassigned inside the poly(A) tail due to an adenine in the reference and these cases were corrected if found. Sequences with at least 50 reads in the DMSO libraries were kept to avoid noise from lowly abundant RNA sequences in the IVT pool. 158,298 L3 variants and 103,018 SVL variants were left after this read depth filtering step. A pseudocount of 1 was then added to variants in L3 2.5 μM, L3 12.5 μM, and SVL 12.5 μM due to drug-mediated drop out of high abundance variants in the DMSO libraries. This pseudocount avoids having undefined drug sensitivities in later steps that would be introduced by log(0). Each variant that passed these checks were counted and converted to a percentage within each RNA-seq library to account for sequencing depth by dividing by the total number of kept reads.

Drug sensitivity for each variant in each dose of Compound 2 was defined as the log ratio of normalized reads from drug-treated RNA-seq divided by the normalized reads from DMSO-treated RNA-seq. Within a given drug dose, sequences with higher log ratios are more resistant than those sequences with lower log ratios.

Minimum free energy folding of IVT RNA's

Minimum free energy (MFE) predictions were done with RNAStructure version 6.4's *Fold* (Reuter and Mathews 2010) and the entire IVT RNA sequence was used. ΔG of each MFE were determined with RNAStructure's *efn2* with command line argument `--simple`. ΔG's from the top 10,000 resistant and sensitive sequences were compared (Quantification and statistical analysis).

C3PO machine learning architecture and training

The architecture chosen for predicting drug sensitivity is based on a previously published 3-layer convolutional neural network (CNN) designed for predicting polysome profiles.(Sample et al. 2019) The model takes in 25 nt one-hot encoded sequences followed by:

First convolution layer: 120 filters (8 × 4), batch normalization, ReLU activation, zero-padding to maintain the same length input and output, and 0% dropout.

Second convolution layer: 120 filters (8 × 1), batch normalization, ReLU activation, zero-padding to maintain the same length input and output, and 0% dropout.

Third convolution layer: 120 filters (8 × 1), batch normalization, ReLU activation, zero-padding to maintain the same length input and output, and 0% dropout.

Dense layer: 80 nodes, batch normalization, ReLU activation and 10% dropout.

Output layer: 3 linear outputs.

The Adam optimizer (Kingma and Ba 2017) was used for model fitting with a mean squared error loss function, batch size of 64, and sample weights based on DMSO read depth.

Sequences were assembled into test and training sets to mix highly covered variants from both RNA contexts (L3 and SVL) into the test and training sets. Within each RNA context, sequences were ordered by DMSO read depth, then split based on the sequences' number in this ordering into odd and even lists, and then the odd and even lists were concatenated together. This odd/even splitting is to include high coverage sequences in both the training and test sets. Finally, the L3 and SVL sequences were interleaved to make an even coverage between RNA contexts in the test set. The top 4,120 sequences were used as the test set with the remaining sequences used as the training set. The 4,120 test set size was chosen because it reflects 2% of the variant space in SVL which contains less variants than L3.

Ten iterations of training with 6 epochs were conducted to account for slight variations in model performance due to stochasticity in the training algorithm. Performance between iterations were evaluated by the square of Pearson's r (R^2) between measured and predicted Compound 2 sensitivity in the test sequences. The best performing iteration was kept and used in further analyses.

Exploring additional machine learning architectures and training

Additional deep learning and training pipelines were explored based on CNN's and Dilated Residual Networks. With the C3PO CNN architecture, training was done with 4-8 epochs and these number of epochs performed relatively similarly on the test set. We also explored using a validation set (4,120 sequences) derived from the training set to determine an early stopping criterion for the number of epochs trained, and this performed similarly to the models trained with a preset number of epochs. Among the three drug doses predictions, 12.5 μM predictions performed better leading us to train models for only this dose as well. However, 12.5 μM prediction performance between the three dose predictions and the one dose prediction were negligibly different so we used the model with three dose predictions.

Hyperband training (Li et al. 2017) with the CNN architecture was also performed to ascertain potential optimal hyperparameter values. Hyperparameters were allowed to range from 1-5 1D convolutional layers with ReLU activation and batch normalization; 8-140 (step 16) number of filters; followed by pooling choices of average, max, or none; and dropout rates of 0-0.5 (step 0.1). These convolutional layer(s) are followed by a Flatten layer, and 1-3 dense layers. Each dense layer can be of size 20-200 (step 20) with ReLU activation, batch normalization, and dropout rates of 0-0.5 (step 0.1). Learning rate parameters were also allowed to range between $1x10^{-5}$-$1x10^{-1}$. Training was allowed to stop early based on the validation set's mean squared error and a minimum delta of 0.001 and patience of 5 epochs. Hyperband training was done with an output layer for all three drug doses, and hyperband training was also tried with an output layer for only predicting 12.5 μM Compound 2 resistance.

Due to the improvement of APARENT2 (Linder et al. 2022) which is a residual neural network (ResNet) over APARENT which is a CNN, we also tried an architecture similar to

APARENT2 with our task on predicting Compound 2 sensitivity. We tested the residual neural network architecture with predicting both Compound 2 sensitivity and cleavage patterns with the hypothesis that learning sequence features that affect cleavage site usage would improve the Compound 2 sensitivity predictions. Input to the residual network is a one-hot encoded 25 nt sequence which is the same as our CNN models and is followed by 20 residual blocks where each block contains 2 layers of dilated convolutions and a skip connection. More specifically, there are 5 residual groups where each residual group contains 4 residual blocks with 32 channels and convolutional filters of size 3. Each residual block is encoded the same as APARENT2 where each blocks has two one-dimensional convolutional layers with batch-normalization, ReLU activation, and a filter dilation rate. There are additional skip connections from between each residual group to the last convolutional layer and produces a vector of length 26, s(x). The 26th position is for all cuts found at positions not found the 25 nt randomized region. For training and accounting for any background sequence biases, a boolean is passed to indicate whether the data point is from the L3 or SVL background which is multiplied with a position-specific weight matrix and linearly combined with s(x). We also kept APARENT2's random shifting of the input sequence and cleavage distribution during training to force the network to not simply learn the designed expected cleavage position in each library. These scores containing library-specific information are sent to four different linear dense layers for separate predictions of cleavage profiles of all four drug doses and softmax transformation is applied to each. For Compound 2 sensitivity prediction, s(x) undergoes average pooling, and the library indicator is concatenated before a linear dense layer for final output. KL-divergence is used as the loss function for cleavage profiles and mean squared error for Compound 2 sensitivities. Total loss is a weighted average of half from Compound 2 sensitivities, and the other half split evenly between the four cleavage profiles. The ResNet was trained with Keras's implementation of the Adam optimizer, batch size of 64, stopping criteria based on a validation set (4,120 sequences) derived from the training set.

We first tried 1, 2, 4, 2, and 1 as dilation rates for the 5 residual groups and performed similarly to previously trained CNNs with R^2 values of 0.232, 0.541, and. 0.681 for the three Compound 2 doses but did not outperform C3PO (Table 3.2). We also tried lower dilation rates of 1, 2, 2, 2, and 1 as well as 1, 1, 1, 1, and 1 which performed worse. Using the dilation rates 1, 2, 4, 2, and 1, we trained for exactly 7 epochs and did not find improved performance. We also increased the cleavage profile length to 27 to separately model cuts found at positions greater than the 25 nt randomized region in position 26, and position 27 is filled when a sequence is not found at a given Compound 2 dose (i.e. sensitive sequences that drop out at higher Compound 2 doses). This led to R^2 values of 0.229, 0.55, and. 0.686 for the three Compound 2 doses which also did not outperform C3PO (Table 3.2). Finally, we increased the weight of Compound 2 sensitivity predictions to 75% of the total loss which did not lead to better performance than C3PO.

Convolutional layers 1 and 2 activation analysis

Convolutional layers 1 and 2 were analyzed similarly to a previously published analysis of a CNN that predicts alternative polyadenylation (APARENT) (Bogard et al. 2019). In brief, every filter in both convolutional layers were correlated with predictions of drug sensitivity at the 12.5 μM dose. The top 5,000 input sequences from the training set that achieved maximal filter activation were put into a position weight matrix and used to generate position-aware consensus sequence logos (Alipanahi et al. 2015). Pearson's r plots of each filter's activations with predicted 12.5 μM Compound 2 sensitivity at each position are plotted below these filter-specific sequence logos. Layer 1 filters are 8 positions wide, and layer 2 filters are 15 positions wide. Note that the convolutional layers in C3PO contain even zero-padding to maintain an input/output size of 25. The padding should be accounted for when analyzing the filters' Pearson r plots. For example in layer 1, the sequences are padded with 4 0's on both the left and right.

APARENT2 predictions and comparisons

APARENT2 predictions of logodds of cleavage at expected cleavage position versus elsewhere were done on all MPRA sequences, centered at their expected cut site which is the expected format of APARENT2. Predictions with read depth of at least 150 in the Input libraries were kept for further analysis. APARENT2 predictions were compared against the logodds of expected cleaved DMSO read counts and Input read counts which estimates the *in vitro* cleavage efficiency. Additionally, APARENT2 predictions were compared against the logodds of expected cleaved 12.5 μM Compound 2 read counts and Input read counts which estimates the *in vitro* drug resistance.

4sU-seq

HepG2 cells were treated with DMSO or 20 μM JTE-607 (Tocris) for 3 hours at 37˚C. 500 μM 4sU (Sigma) was then added to the DMSO/JTE-607 containing media and cells were incubated at 37˚C for one additional hour. After incubation, cells were lysed in Trizol (Invitrogen), and total RNA was extracted following the manufacturer's protocol. 4sU RNA enrichment and library preparation were done as previously described with minor modifications (Wang et al. 2020). Briefly, 50 µg total RNA was used as the starting material and biotinylated with biotin-HPDP (Thermo). 4sU labeled and biotinylated RNA was enriched with streptavidin beads by rotating at room temperature for 1.5 hours, eluted with 100 mM DTT, and further purified by phenol chloroform extraction.

PAS-seq

HepG2 cells were treated with DMSO or 20 µM JTE-607 (Tocris) for 4 hours at 37˚C and total RNA was extracted by Trizol (Invitrogen). 10 µg of total RNA was used to prepare PAS-seq libraries as previously described (Wang et al. 2021). Briefly, 10 µg total RNA was fragmented using fragmentation buffer (Thermo) and reverse transcribed by SuperScript III (Thermo). The cDNA was circularized by Circligase (Lucigen) and then re-linearized by BamHI. The digested DNA was then PCR amplified and a ~200 bp region was gel extracted and sequenced. During the execution of this study, we have further optimized the library preparation steps of the protocol and for the most updated PAS-seq protocol please refer to PAS-seq 2 (Yoon et al. 2021).

4sU-seq and PAS-seq data analysis

For 4sU-seq, reads were mapped to human hg19 using STAR (Dobin et al. 2013) and bigwig files were generated using deepTools (Ramírez et al. 2014). For PAS-seq, reads without a poly(A) tail (fewer than 15 consecutive A's) were removed. The polyA tail sequence and linker sequence was trimmed from remaining reads before mapping. The trimmed reads were mapped to human hg19 using STAR (Dobin et al. 2013) as done for 4sU-seq except that the EndToEnd parameter was used only for PAS-seq. The resulting bam output file was converted to a bed file using BEDTools (Quinlan and Hall 2010). Reads that may have been due to internal priming (reads where there were 6 consecutive A's within 10 nucleotides downstream of the PAS, or 7 A's out of 10 nucleotides downstream of the PAS) were removed. The resulting bed file was then converted back to a bam file using BEDTools (Quinlan and Hall 2010). The location of the 3′ end of each read was extracted using BEDTools (Quinlan and Hall 2010) and was then compared to the location of all annotated PAS within PolyA_DB (Wang et al. 2018) to retrieve read counts for each PAS. APA analysis was performed by edgeR (Robinson et al. 2010).

Quantification and statistical analysis

Pearson's r and R^2 (square of Pearson's r) are used in Figures 3.1, 3.3, 3.5, 3.10, 3.4, 3.7, 3.11 and 3.12 and related text as well as SI Models Table. Potential inequality of the top 10,000 resistant and sensitive sequences' MFE ΔG's were tested with a two-sided t-test with unequal variance. 6-mers in the top 10,000 resistant and sensitive sequences were found to be significant by a binomial test with a null hypothesis of probability of success = 0.25^6 and alternative hypothesis of > 0.25^6. p-value threshold was adjusted by the number of possible k-mers, 4^6, and thus significant 6-mers must have p-values $\leq \frac{0.05}{4^6}$.

CHAPTER 4

Perspectives

mRNA 3′ processing, also known as the cleavage and polyadenylation of pre-mRNA, is a complex co-transcriptional event that plays a crucial role in the gene expression pathway. The identification of the components in the mammalian mRNA 3' processing complex, along with recent advancements in structural determination through cryo-electron microscopy, have generated a more comprehensive understanding of protein-protein and protein-RNA interactions within the mRNA 3′ processing complex. These interactions are integral to elucidating the regulatory processes governing mRNA 3′ processing within cells.

In recent years, numerous viral infections and small molecule inhibitors have been found to interfere with mRNA 3′ processing. As a result, mRNA 3′ processing has emerged as a potentially promising target for the development of novel antiviral and anticancer compounds. However, the relative lack of knowledge regarding the regulatory mechanisms of mRNA 3′ processing poses a significant challenge for drug development.

This work aimed to bridge the gap in our mechanistic understanding of mRNA 3′ processing regulation by employing both viral infection and small molecule inhibitors as research tools. By gaining a deeper understanding into the regulatory mechanism of mRNA 3′ processing, we hope to facilitate the development of novel and effective therapeutic interventions.

Integrating multiple high-throughput sequencing approaches including nascent RNA sequencing (4sU-seq), poly(A) site sequencing (PAS-seq), fractionation RNA sequencing (RNA-seq), and ribosome profiling (Ribo-seq), Chapter 2 elucidated the extent, mechanism, and functional impact of HSV-1 infection on the host APA. HSV-1 induced widespread and dynamic APA changes in infected cells, with a majority of these changes involved intronic PAS. At the same time, HSV-1 infection also induced DoTT. We further provided evidence that the viral immediate early protein ICP27 is necessary but not sufficient for HSV-1-induced APA changes. While the transcripts cleaved and polyadenylated at intronic PAS are exported and at least some of them are translated, the long readthrough transcripts are not exported. In addition to HSV-1, multiple other viruses and cellular stress modulate APA. It will be important for future research to evaluate whether a common mechanism underlies virus and stress-mediated regulation of mRNA 3′ processing.

In Chapter 3 of this work, we used a combination of in vitro biochemistry, machine learning, and high-throughput sequencing approaches to uncover the sequence-specific regulation of mRNA 3′ processing by the anti-cancer compound JTE-607. The 25 nt cleavage site region immediately following AAUAAA is the key sequence determinant for JTE-607 sensitivity. Our machine learning program, C3PO, can predict the JTE-607 sensitivity for any given PAS both in vitro and in cells with great accuracy.

A particularly novel and exciting aspect of this study, which merges the work presented in both Chapters 2 and 3, is the potential for mRNA 3′ processing to serve as a novel therapeutic target in the treatment of various diseases, such as cancers and viral infections. This possibility opens up new avenues for the development of innovative medical interventions.

One such example is the potential of utilizing the manipulation of mRNA 3′ processing by JTE-607 in HIV latency reversal. The HIV genome contains two identical long terminal repeats (LTRs) that are present at the 5′ and 3′ end. Each of these LTRs contain a PAS, which is also the same in the sequence content. Usage of the 5′ LTR PAS results in production of short HIV transcripts that do not encode HIV proteins. This may be important for the establishment and maintenance of HIV latency, where there is no productive viral gene expression and replication. In contrast, usage of the 3′ LTR PAS leads to production of full-length HIV transcript, leading to viral gene expression. If the PAS usage in cells that are latently infected with HIV can be shifted from the 5′ LTR PAS to the 3′ LTR PAS, this will activate the production of full-length HIV genome and viral replication, which in turn leads to the release of virus from the infected cells and detection of the virus by the host immune system and viral elimination. As JTE-607 treatment in cell culture predominately induced proximal to distal APA changes, it will be interesting to test the effect of JTE-607 on HIV PAS selection. Moreover, despite the acceptable efficiency of current latency reversing agents (LRAs), a variety of cytokines are produced after LRA treatment, which may be detrimental to the host. As JTE-607 is also a potent cytokine production inhibitor, JTE-607 and LRA treatment on cells latently infected with HIV may exhibit a combinatorial effect: improving the efficiency of latency reversal and at the same time, decreasing cytokine production and reducing the adverse effect of LRAs. Future research is needed to understand the relationship between JTE-607 and HIV latency reversal.

In conclusion, this study has revealed some of the mechanisms governing mRNA 3′ processing regulation during HSV-1 infection and in the presence of the small molecule inhibitor JTE-607. Additionally, the work presented here has highlighted the potential for targeting mRNA 3′ processing as a therapeutic strategy to combat a range of diseases and viral infections, which should be beneficial in many different fields, such as cancer biology and virology.

REFERENCES

Abmayr SM, Yao T, Parmely T, Workman JL. 2006. Preparation of nuclear and cytoplasmic extracts from mammalian cells. *Curr Protoc Mol Biol* **Chapter 12**: Unit 12.1.

Alahmari AA, Chaubey AH, Tisdale AA, Schwarz CD, Cornwell AC, Maraszek KE, Paterson EJ, Kim M, Venkat S, Gomez EC, et al. 2022. CPSF3 inhibition blocks pancreatic cancer cell proliferation through disruption of core histone processing. 2022.05.09.491230. https://www.biorxiv.org/content/10.1101/2022.05.09.491230v1 (Accessed November 3, 2022).

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–838.

Aravind L. 1999. An evolutionary classification of the metallo-beta-lactamase fold proteins. *In Silico Biol* **1**: 69–91.

Ay S, Di Nunzio F. 2023. HIV-Induced CPSF6 Condensates. *Journal of Molecular Biology* 168094.

Batra R, Stark TJ, Clark AE, Belzile J-P, Wheeler EC, Yee BA, Huang H, Gelboin-Burkhart C, Huelga SC, Aigner S, et al. 2016. RNA-binding protein CPEB1 remodels host and viral RNA landscapes. *Nat Struct Mol Biol* **23**: 1101–1110.

Bauer DLV, Tellier M, Martínez-Alonso M, Nojima T, Proudfoot NJ, Murphy S, Fodor E. 2018. Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription. *Cell Rep* **23**: 2119-2129.e3.

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10**: 1001–1010.

Begolo D, Vincent IM, Giordani F, Pöhner I, Witty MJ, Rowan TG, Bengaly Z, Gillingwater K, Freund Y, Wade RC, et al. 2018. The trypanocidal benzoxaborole AN7973 inhibits trypanosome mRNA processing. *PLoS Pathogens* **14**. https://pubmed.ncbi.nlm.nih.gov/30252911/ (Accessed February 15, 2021).

Bejarano DA, Peng K, Laketa V, Börner K, Jost KL, Lucic B, Glass B, Lusic M, Müller B, Kräusslich H-G. 2019. HIV-1 nuclear import in macrophages is regulated by CPSF6-capsid interactions at the nuclear pore complex eds. W.I. Sundquist, W. Li, A. Engelman, and G.J. Towers. *eLife* **8**: e41800.

Bentley DL. 2005. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* **17**: 251–256.

Bienroth S, Wahle E, Suter-Crazzolara C, Keller W. 1991. Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. *J Biol Chem* **266**: 19768–19776.

Bogard N, Linder J, Rosenberg AB, Seelig G. 2019. A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell* **178**: 91-106.e23.

Boreikaite V, Elliott TS, Chin JW, Passmore LA. 2022. RBBP6 activates the pre-mRNA 3′ end processing machinery in humans. *Genes Dev* **36**: 210–224.

Borozdenkova S, Mant TGK, Allen E, Pu K, Hoshino S, Jurcevic S. 2011. Effects of a cytokine inhibitor, JTE-607, on the response to endotoxin in healthy human volunteers. *Int Immunopharmacol* **11**: 1837–1843.

Bushnell B, Rood J, Singer E. 2017. BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE* **12**: e0185056.

Callebaut I, Moshous D, Mornon J-P, de Villartay J-P. 2002. Metallo-β-lactamase fold within nucleic acids processing enzymes: the β-CASP family. *Nucleic Acids Res* **30**: 3592–3601.

Calzado MA, Sancho R, Muñoz E. 2004. Human immunodeficiency virus type 1 Tat increases the expression of cleavage and polyadenylation specificity factor 73-kilodalton subunit modulating cellular and viral expression. *J Virol* **78**: 6846–6854.

Chan S, Choi EA, Shi Y. 2011. Pre-mRNA 3'-end processing complex assembly and function. *Wiley Interdiscip Rev RNA* **2**: 321–335.

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev* **28**: 2370–2380.

Chen C-C, Li B, Millman SE, Chen C, Li X, Morris JP, Mayle A, Ho Y-J, Loizou E, Liu H, et al. 2020. Vitamin B6 Addiction in Acute Myeloid Leukemia. *Cancer Cell* **37**: 71-84.e7.

Chen F, MacDonald CC, Wilusz J. 1995. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res* **23**: 2614–2620.

Colgan DF, Manley JL. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11**: 2755–2766.

Colgan DF, Murthy KG, Prives C, Manley JL. 1996. Cell-cycle related regulation of poly(A) polymerase by phosphorylation. *Nature* **384**: 282–285.

Cortazar MA, Sheridan RM, Erickson B, Fong N, Glover-Cutter K, Brannan K, Bentley DL. 2019. Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Molecular Cell* **76**: 896-908.e4.

Coseno M, Martin G, Berger C, Gilmartin G, Keller W, Doublié S. 2008. Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic Acids Res* **36**: 3474–3483.

Dai-Ju JQ, Li L, Johnson LA, Sandri-Goldin RM. 2006. ICP27 Interacts with the C-Terminal Domain of RNA Polymerase II and Facilitates Its Recruitment to Herpes Simplex Virus 1 Transcription Sites, Where It Undergoes Proteasomal Degradation during Infection. *Journal of Virology* **80**: 3567–3581.

Dass B, Tardif S, Park JY, Tian B, Weitlauf HM, Hess RA, Carnes K, Griswold MD, Small CL, MacDonald CC. 2007. Loss of polyadenylation protein τCstF-64 causes spermatogenic defects and male infertility. *Proceedings of the National Academy of Sciences* **104**: 20374–20379.

Davis R, Shi Y. 2014. The polyadenylation code: a unified model for the regulation of mRNA alternative polyadenylation. *J Zhejiang Univ Sci B* **15**: 429–437.

de la Vega L, Sánchez-Duffhues G, Fresno M, Schmitz ML, Muñoz E, Calzado MA. 2007. The 73 kDa subunit of the CPSF complex binds to the HIV-1 LTR promoter and functions as a negative regulatory factor that is inhibited by the HIV-1 Tat protein. *J Mol Biol* **372**: 317–330.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173–1183.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell* **43**: 853–866.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dubbury SJ, Boutz PL, Sharp PA. 2018. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature* **564**: 141–145.

Eaton JD, Francis L, Davidson L, West S. 2020. A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes Dev* **34**: 132–145.

Eaton JD, West S. 2020. Termination of Transcription by RNA Polymerase II: BOOM! *Trends in Genetics* **36**: 664–675.

Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, Stevanovic S, Zimmer R, Dölken L. 2018. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* **15**: 363–366.

Fraser KA, Rice SA. 2007. Herpes Simplex Virus Immediate-Early Protein ICP22 Triggers Loss of Serine 2-Phosphorylated RNA Polymerase II. *Journal of Virology* **81**: 5091–5101.

Galluzzi L, Yamazaki T, Kroemer G. 2018. Linking cellular stress responses to systemic homeostasis. *Nat Rev Mol Cell Biol* **19**: 731–745.

Gutierrez PA, Baughman K, Sun Y, Tong L. 2021. A real-time fluorescence assay for CPSF73, the nuclease for pre-mRNA 3'-end processing. *RNA* **27**: 1148–1154.

Hennig T, Michalski M, Rutkowski AJ, Djakovic L, Whisnant AW, Friedl M-S, Jha BA, Baptista MAP, L'Hernault A, Erhard F, et al. 2018. HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes ed. R.F. Kalejta. *PLoS Pathog* **14**: e1006954.

Hill CH, Boreikaitė V, Kumar A, Casañal A, Kubík P, Degliesposti G, Maslen S, Mariani A, von Loeffelholz O, Girbig M, et al. 2019. Activation of the Endonuclease that Defines mRNA 3′ Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Molecular Cell* **73**: 1217-1231.e11.

Huang K-L, Jee D, Stein CB, Elrod ND, Henriques T, Mascibroda LG, Baillat D, Russell WK, Adelman K, Wagner EJ. 2020. Integrator Recruits Protein Phosphatase 2A to Prevent Pause Release and Facilitate Transcription Termination. *Molecular Cell* **80**: 345-358.e9.

Jia X, Yuan S, Wang Y, Fu Y, Ge Y, Ge Y, Lan X, Feng Y, Qiu F, Li P, et al. 2017. The role of alternative polyadenylation in the antiviral innate immune response. *Nat Commun* **8**: 14605.

Jian MY, Koizumi T, Tsushima K, Kubo K. 2004. JTE-607, a cytokine release blocker, attenuates acid aspiration-induced lung injury in rats. *European Journal of Pharmacology* **488**: 231–238.

Johnson SA, Cubberley G, Bentley DL. 2009. Cotranscriptional recruitment of the mRNA export factor Yra1 by direct interaction with the 3' end processing factor Pcf11. *Mol Cell* **33**: 215–226.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664–668.

Kakegawa J, Sakane N, Suzuki K, Yoshida T. 2019. JTE-607, a multiple cytokine production inhibitor, targets CPSF3 and inhibits pre-mRNA processing. *Biochemical and Biophysical Research Communications* **518**: 32–37.

Kakutani M, Takeuchi K, Waga I, Iwamura H, Wakitani K. 1999. JTE-607, a novel inflammatory cytokine synthesis inhibitor without immunosuppression, protects from endotoxin shock in mice. *Inflammation Research* **48**: 461–468.

Kamieniarz-Gdula K, Gdula MR, Panser K, Nojima T, Monks J, Wiśniewski JR, Riepsaame J, Brockdorff N, Pauli A, Proudfoot NJ. 2019. Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol Cell* **74**: 158-172.e9.

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**: 616–626.

Kecman T, Kuś K, Heo D-H, Duckett K, Birot A, Liberatori S, Mohammed S, Geis-Asteggiante L, Robinson CV, Vasiljeva L. 2018. Elongation/Termination Factor Exchange Mediated by PP1 Phosphatase Orchestrates Transcription Termination. *Cell Reports* **25**: 259-269.e5.

Kingma DP, Ba J. 2017. *Adam: A Method for Stochastic Optimization*. arXiv http://arxiv.org/abs/1412.6980 (Accessed September 5, 2022).

Krajewska M, Dries R, Grassetti AV, Dust S, Gao Y, Huang H, Sharma B, Day DS, Kwiatkowski N, Pomaville M, et al. 2019. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat Commun* **10**: 1757.

Lackford B, Yao C, Charles GM, Weng L, Zheng X, Choi E-A, Xie X, Wan J, Xing Y, Freudenberg JM, et al. 2014. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J* **33**: 878–889.

Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. 2017. Hyperband: a novel bandit-based approach to hyperparameter optimization. *J Mach Learn Res* **18**: 6765–6816.

Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, Coller J, Pasquinelli AE. 2017. Short Poly(A) Tails are a Conserved Feature of Highly Expressed Genes. *Nat Struct Mol Biol* **24**: 1057–1063.

Lin M-H, Jensen MK, Elrod ND, Huang K-L, Welle KA, Wagner EJ, Tong L. 2022. Inositol hexakisphosphate is required for Integrator function. *Nat Commun* **13**: 5742.

Linder J, Koplik SE, Kundaje A, Seelig G. 2022. Deciphering the impact of genetic variation on human polyadenylation using APARENT2. *Genome Biol* **23**: 232.

Liu H, Heller-Trulli D, Moore CL. 2022. Targeting the mRNA endonuclease CPSF73 inhibits breast cancer cell migration, invasion, and self-renewal. *iScience* **25**: 104804.

Liu X, Hoque M, Larochelle M, Lemay J-F, Yurko N, Manley JL, Bachand F, Tian B. 2017. Comparative analysis of alternative polyadenylation in S. cerevisiae and S. pombe. *Genome Res* **27**: 1685–1695.

Ludwig A, Hengel H. 2009. Vesicular Stomatitis Virus Infection. In *Encyclopedia of Molecular Mechanisms of Disease* (ed. F. Lang), pp. 2204–2205, Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-540-29676-8_1841 (Accessed May 4, 2023).

Lukiw WJ, Bazan NG. 1997. Cyclooxygenase 2 RNA message abundance, stability, and hypervariability in sporadic Alzheimer neocortex. *J Neurosci Res* **50**: 937–945.

MacDonald CC, Redondo J-L. 2002. Reexamining the polyadenylation signal: were we wrong about AAUAAA? *Mol Cell Endocrinol* **190**: 1–8.

Mandel CR, Bai Y, Tong L. 2007. Protein factors in pre-mRNA 3′-end processing. *Cellular and Molecular Life Sciences 2007 65:7* **65**: 1099–1122.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444**: 953–956.

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide Analysis of Pre-mRNA 3' End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3' UTR Length. *Cell Reports* **1**: 753–763.

Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu A-B, Li W, Wagner EJ. 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**: 412–416.

Mayr C. 2016. Evolution and Biological Roles of Alternative 3′UTRs. *Trends in Cell Biology* **26**: 227–237.

McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, Patterson SD, Wickens M, Bentley DL. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**: 357–361.

Mitschka S, Mayr C. 2022. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol*.

Moore CL, Chen J, Whoriskey J. 1988. Two proteins crosslinked to RNA containing the adenovirus L3 poly(A) site require the AAUAAA sequence for binding. *The EMBO Journal* **7**: 3159–3169.

Moore MJ, Proudfoot NJ. 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**: 688–700.

Murthy KG, Manley JL. 1992. Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J Biol Chem* **267**: 14804–14811.

Murthy KG, Manley JL. 1995. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev* **9**: 2672–2683.

Nemeroff ME, Barabino SM, Li Y, Keller W, Krug RM. 1998. Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3'end formation of cellular pre-mRNAs. *Mol Cell* **1**: 991–1000.

Nguyen-Tran H, Messacar K. 2022. Preventing enterovirus A71 disease: another promising vaccine for children. *The Lancet* **399**: 1671–1673.

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**: 526–540.

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143**: 1018–1029.

Pai AA, Baharian G, Sabourin AP, Brinkworth JF, Nédélec Y, Foley JW, Grenier J-C, Siddle KJ, Dumaine A, Yotova V, et al. 2016. Widespread Shortening of 3' Untranslated Regions and Increased Exon Inclusion Are Evolutionarily Conserved Features of Innate Immune Responses to Infection. *PLOS Genetics* **12**: e1006338.

Palencia A, Bougdour A, Brenier-Pinchart M, Touquet B, Bertini R, Sensi C, Gay G, Vollaire J, Josserand V, Easom E, et al. 2017. Targeting *Toxoplasma gondii* CPSF3 as a new approach to control toxoplasmosis. *EMBO Molecular Medicine* **9**: 385–394.

Proudfoot N. 2004. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**: 272–278.

Proudfoot NJ. 2016. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**: aad9926–aad9926.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: W187-191.

Reuter JS, Mathews DH. 2010. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**: 129.

Richard P, Manley JL. 2009. Transcription termination by nuclear RNA polymerases. *Genes Dev* **23**: 1247–1269.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Ross NT, Lohmann F, Carbonneau S, Fazal A, Weihofen WA, Gleim S, Salcius M, Sigoillot F, Henault M, Carl SH, et al. 2020. CPSF3-dependent pre-mRNA processing as a druggable node in AML and Ewing's sarcoma. *Nature Chemical Biology* **16**: 50–59.

Rüegsegger U, Beyer K, Keller W. 1996. Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* **271**: 6107–6113.

Rüegsegger U, Blank D, Keller W. 1998. Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell* **1**: 243–253.

Ruepp M-D, Schweingruber C, Kleinschmidt N, Schümperli D. 2011. Interactions of CstF-64, CstF-77, and symplekin: Implications on localisation and function ed. A.G. Matera. *MBoC* **22**: 91–104.

Rutkowski AJ, Erhard F, L'Hernault A, Bonfert T, Schilhabel M, Crump C, Rosenstiel P, Efstathiou S, Zimmer R, Friedel CC, et al. 2015. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun* **6**: 7126.

Ryner LC, Takagaki Y, Manley JL. 1989. Sequences downstream of AAUAAA signals affect pre-mRNA cleavage and polyadenylation in vitro both directly and indirectly. *Mol Cell Biol* **9**: 1759–1771.

Ryugo M, Sawa Y, Ono M, Miyamoto Y, Aleshin AN, Matsuda H. 2004. Pharmacologic preconditioning of JTE-607, a novel cytokine inhibitor, attenuates ischemia-reperfusion injury in the myocardium. *J Thorac Cardiovasc Surg* **127**: 1723–1727.

Sadowski M, Dichtl B, Hübner W, Keller W. 2003. Independent functions of yeast Pcf11p in pre-mRNA 3′ end processing and in transcription termination. *The EMBO Journal* **22**: 2167–2177.

Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, Seelig G. 2019. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* **37**: 803–809.

Sandri-Goldin RM. 2011. The many roles of the highly interactive HSV protein ICP27, a key regulator of infection. *Future Microbiology* **6**: 1261–1277.

Sasaki J, Fujishima S, Iwamura H, Wakitani K, Aiso S, Aikawa N. 2003. Prior burn insult induces lethal acute lung injury in endotoxemic mice: Effects of cytokine inhibition. *American Journal of Physiology - Lung Cellular and Molecular Physiology* **284**. https://pubmed.ncbi.nlm.nih.gov/12388363/ (Accessed February 15, 2021).

Schmidt M, Kluge F, Sandmeir F, Kühn U, Schäfer P, Tüting C, Ihling C, Conti E, Wahle E. 2022. Reconstitution of 3′ end processing of mammalian pre-mRNA reveals a central role of RBBP6. *Genes Dev* **36**: 195–209.

Schwich OD, Blümel N, Keller M, Wegener M, Setty ST, Brunstein ME, Poser I, Mozos IRDL, Suess B, Münch C, et al. 2021. SRSF3 and SRSF7 modulate 3'UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. *Genome biology* **22**. https://pubmed.ncbi.nlm.nih.gov/33706811/ (Accessed May 16, 2022).

Sekulovich RE, Leary K, Sandri-Goldin RM. 1988. The herpes simplex virus type 1 alpha protein ICP27 can act as a trans-repressor or a trans-activator in combination with ICP4 and ICP0. *J Virol* **62**: 4510–4522.

Sheets MD, Ogg SC, Wickens MP. 1990. Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* **18**: 5799–5805.

Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761–772.

Shi Y. 2012. Alternative polyadenylation: new insights from global analyses. *RNA* **18**: 2105–2117.

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR, Frank J, Manley JL. 2009. Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. *Molecular Cell* **33**: 365–376.

Shi Y, Manley JLJL. 2015. The end of the message: multiple protein--RNA interactions define the mRNA polyadenylation site. *Genes & development* **29**: 889–897.

Smith IL, Hardwicke MA, Sandri-Goldin RM. 1992. Evidence that the herpes simplex virus immediate early protein ICP27 acts post-transcriptionally during infection to regulate gene expression. *Virology* **186**: 74–86.

Soles LV, Shi Y. 2021. Crosstalk Between mRNA 3'-End Processing and Epigenetics. *Front Genet* **12**: 637705.

Sonoiki E, Ng CL, Lee MCS, Guo D, Zhang YK, Zhou Y, Alley MRK, Ahyong V, Sanz LM, Lafuente-Monasterio MJ, et al. 2017. A potent antimalarial benzoxaborole targets a Plasmodium falciparum cleavage and polyadenylation specificity factor homologue. *Nature Communications* **8**: 1–11.

Sowd GA, Serrao E, Wang H, Wang W, Fadel HJ, Poeschla EM, Engelman AN. 2016. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* **113**: E1054-1063.

Sun Y, Zhang Y, Aik WS, Yang XC, Marzluff WF, Walz T, Dominski Z, Tong L. 2020. Structure of an active human histone pre-mRNA 3′-end processing machinery. *Science* **367**: 700–703.

Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci USA* **115**. https://pnas.org/doi/full/10.1073/pnas.1718723115 (Accessed August 2, 2022).

Swale C, Bougdour A, Gnahoui-David A, Tottey J, Georgeault S, Laurent F, Palencia A, Hakimi M-A. 2019. Metal-captured inhibition of pre-mRNA processing activity by CPSF3 controls *Cryptosporidium* infection. *Sci Transl Med* **11**: eaax7161.

Takagaki Y, Manley JL. 2000. Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol Cell Biol* **20**: 1515–1525.

Takagaki Y, Manley JL. 1997. RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol* **17**: 3907–3914.

Takagaki Y, Seipelt RL, Peterson ML, Manley JL. 1996. The Polyadenylation Factor CstF-64 Regulates Alternative Processing of IgM Heavy Chain Pre-mRNA during B Cell Differentiation. *Cell* **87**: 941–952.

Tang S, Patel A, Krause PR. 2016. Herpes simplex virus ICP27 regulates alternative pre-mRNA polyadenylation and splicing in a sequence-dependent manner. *Proceedings of the National Academy of Sciences* **113**: 12256–12261.

Tian B, Manley JL. 2016. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology* **18**: 18–30.

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors HHS Public Access. *Nat Rev Mol Cell Biol* **18**: 18–30.

Twu KY, Noah DL, Rao P, Kuo R-L, Krug RM. 2006. The CPSF30 Binding Site on the NS1A Protein of Influenza A Virus Is a Potential Antiviral Target. *J Virol* **80**: 3957–3965.

Uesato N, Fukui K, Maruhashi J, Tojo A, Tajima N. 2006. JTE-607, a multiple cytokine production inhibitor, ameliorates disease in a SCID mouse xenograft acute myeloid leukemia model. *Experimental Hematology* **34**: 1385–1392.

Vethantham V, Rao N, Manley JL. 2007. Sumoylation modulates the assembly and activity of the pre-mRNA 3’ processing complex. *Mol Cell Biol* **27**: 8848–8858.

Vethantham V, Rao N, Manley JL. 2008. Sumoylation regulates multiple aspects of mammalian poly(A) polymerase function. *Genes Dev* **22**: 499–511.

Vijayakumar A, Park A, Steitz JA. 2022. Modulation of mRNA 3′-End Processing and Transcription Termination in Virus-Infected Cells. *Frontiers in Immunology* **13**. https://www.frontiersin.org/articles/10.3389/fimmu.2022.828665 (Accessed May 3, 2023).

Vilborg A, Passarelli MC, Yario TA, Tycowski KT, Steitz JA. 2015. Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell* **59**: 449–461.

Wall RJ, Rico E, Lukac I, Zuccotto F, Elg S, Gilbert IH, Freund Y, Alley MRK, Field MC, Wyllie S, et al. 2018. Clinical and veterinary trypanocidal benzoxaboroles target CPSF3. *Proceedings of the National Academy of Sciences* **115**: 9616–9621.

Wan R, Bai R, Yan C, Lei J, Shi Y. 2019. Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. *Cell* **177**: 339-351.e13.

Wang R, Nambiar R, Zheng D, Tian B. 2018. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**: D315–D319.

Wang R, Zheng D, Wei L, Ding Q, Tian B. 2019. Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Rep* **26**: 2766-2778.e6.

Wang X, Hennig T, Whisnant AW, Erhard F, Prusty BK, Friedel CC, Forouzmand E, Hu W, Erber L, Chen Y, et al. 2020. Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27. *Nat Commun* **11**: 293.

Wang X, Liu L, Whisnant AW, Hennig T, Djakovic L, Haque N, Bach C, Sandri-Goldin RM, Erhard F, Friedel CC, et al. 2021. Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation ed. N.K. Conrad. *PLoS Genet* **17**: e1009263.

Whitley RJ, Roizman B. 2001. Herpes simplex virus infections. *Lancet* **357**: 1513–1518.

Wu X, Bartel DP. 2017. Widespread Influence of 3′-End Structures on Mammalian mRNA Processing and Stability. *Cell* **169**: 905-917.e11.

Yang Q, Gilmartin GM, Doublié S. 2010. Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3’ processing. *Proc Natl Acad Sci U S A* **107**: 10062–10067.

Yang W, Hsu PL, Yang F, Song J-E, Varani G. 2018. Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3′-end processing signals. *Nucleic Acids Res* **46**: 493–503.

Yao C, Biesinger J, Wan J, Weng L, Xing Y, Xie X, Shi Y. 2012. Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc Natl Acad Sci USA* **109**: 18773–18778.

Yao C, Choi E-A, Weng L, Xie X, Wan J, Xing Y, Moresco JJ, Tu PG, Yates JR, Shi Y. 2013. Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3′ processing. *RNA* **19**: 1781–1790.

Yoon Y, Soles LV, Shi Y. 2021. PAS-seq 2: A fast and sensitive method for global profiling of polyadenylated RNAs. *Methods in enzymology* **655**: 25–35.

York JD, Odom AR, Murphy R, Ives EB, Wente SR. 1999. A phospholipase C-dependent inositol polyphosphate kinase pathway required for efficient messenger RNA export. *Science* **285**: 96–100.

Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3’ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405–445.

Zhao N, Sebastiano V, Moshkina N, Mena N, Hultquist J, Jimenez-Morales D, Ma Y, Rialdi A, Albrecht R, Fenouil R, et al. 2018. Influenza virus infection causes global RNAPII termination defects. *Nat Struct Mol Biol* **25**: 885–893.

Zheng D, Wang R, Ding Q, Wang T, Xie B, Wei L, Zhong Z, Tian B. 2018. Cellular stress alters 3′UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nat Commun* **9**: 2268.

Zheng Y, Schubert HL, Singh PK, Martins LJ, Engelman AN, D’Orso I, Hill CP, Planelles V. 2021. Cleavage and Polyadenylation Specificity Factor 6 Is Required for Efficient HIV-1 Latency Reversal. *mBio* **12**: e0109821.

Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, Sowd GA, Engelman AN, Xie X, Hertel KJ, Shi Y. 2018. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Molecular Cell* **69**: 62-74.e4.

Zorita E, Cuscó P, Filion GJ. 2015. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**: 1913–1919.